UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA FUNDAMENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

MARCIO FELIPE DE OLIVEIRA

OTIMIZAÇÃO QUIMIOMÉTRICA PARA ENSAIOS METABONÔMICOS BASEADOS EM ESPECTROSCOPIA DE RMN: DIAGNÓSTICO E ESTADIAMENTO DE CÂNCER DE PRÓSTATA

Recife
2024

MARCIO FELIPE DE OLIVEIRA

OTIMIZAÇÃO QUIMIOMÉTRICA PARA ENSAIOS METABONÔMICOS BASEADOS EM ESPECTROSCOPIA DE RMN: DIAGNÓSTICO E ESTADIAMENTO DE CÂNCER DE PRÓSTATA

(Tese de Doutorado)

Tese apresentada ao Programa de Pós-Graduação em Química como pré-requisito para obtenção do título de doutor em Química.

Orientador: Prof. Dr. Ricardo Oliveira da Silva

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Oliveira, Marcio Felipe de.
Otimização quimiométrica para ensaios metabonômicos baseados em espectroscopia de RMN: diagnóstico e estadiamento de câncer de próstata / Marcio Felipe de Oliveira. - Recife, 2024.
120f.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-Graduação em Química, 2024.
Orientação: Ricardo Oliveira da Silva.
Inclui referências e apêndices.

1. Metabolômica; 2. Quimiometria; 3. Neoplasia; 4. Soro; 5. Espectroscopia; 6. Pré-processamento de dados. I. Silva, Ricardo Oliveira da. II. Título.

UFPE-Biblioteca Central

AGRADECIMENTOS

Não há como, nesses parágrafos, atender a todos que estiveram presentes e que participaram dessa conquista. Portanto, àqueles que não puderam ser lembrados aqui, saibam que estarão sempre em meus pensamentos.

Agradeço ao meu querido orientador Ricardo Oliveira da Silva, que ajudou imensamente na minha formação como pesquisador e como professor, seja orientando a pesquisa, seja em conversas informais. Meu muitíssimo obrigado.

Agradeço aos companheiros de LabMeQ por sua ajuda, ouvidos e trocas de experiências, um grupo unido, que está sempre disposto a ajudar.

Agradecimentos ao querido professor Dr. Licarion Pinto, pelo apoio, dicas e palavras de conforto, além de, juntamente com o professor Dr. Aderval Luna, me dar a oportunidade de poder terminar essa etapa com um pouco mais de tranquilidade.

Agradecimentos mais que especiais, aos membros da minha família, em especial à minha mãe, guerreira e batalhadora, que me ajudou e com grande esforço me levou a essas vitórias. Às minhas tias Ana Lúcia e Maria Aparecida, que sempre tinham uma palavra de incentivo. A essas só me resta agradecer infinitamente por sua ajuda.

Agradeço à minha namorada, Agna Xavier, que suportou minhas mudanças de humor, preocupações, afastamentos e teve que ouvir sobre muitos problemas de todas as ordens, durante esse maçante período de aproximadamente cinco anos. Agradeço o carinho, afeto e palavras de apoio e compreensão. Grato pelas suas palavras e, muitas vezes pelo seu silêncio. Te amo!

Agradeço aos meus novos e velhos amigos, que têm estado comigo como parte da minha família e estão sempre presentes nos momentos felizes e de grandes dificuldades. Apesar de muitas vezes interiorizar problemas, com bons papos de mesa de bar ou na casa de alguns de vocês, pude externar minhas aflições. A vocês minha eterna gratidão.

Por fim, agradeço ao corpo administrativo da minha *alma mater* Universidade Federal de Pernambuco, e aos técnicos e professores do Departamento de Química Fundamental, exemplo de grande seriedade com a pesquisa não só para Recife, mas também para todo o Brasil. Agradecimento também ao CNPq/MCT, que com seu programa de bolsas beneficiou não só a mim e a outros pesquisadores em todo o Brasil.

“Um passo a frente e você não está mais no mesmo lugar.”
(Chico Science & Nação Zumbi, 1996)

RESUMO

Uma das dificuldades enfrentadas por estudos metabonômicos, usando dados espectrais de ressonância magnética nuclear (RMN) obtidos de biofluidos de humanos, é o pequeno número de amostras frente ao número de variáveis geradas. Em adição a isso, dados da área de saúde também podem ser desbalanceados, o que pode acarretar sobreajuste nos modelos quimiométricos gerados. Dessa forma, técnicas bastante empregadas na aprendizagem de máquina, como a seleção de variáveis, sobreamostragem e regularização podem ser utilizados para mitigar esses efeitos. No presente trabalho, para mitigar o problema de sobreajuste de dois conjuntos de dados metabonômicos baseados em RMN de ^{1}H já publicados, foram testados de forma exaustiva sete técnicas de seleção de variáveis, e cinco métodos de classificação após reamostragem, gerando 43 combinações possíveis. Além disso, para um conjunto de dados inédito, foi realizado o mesmo processo exaustivo, para descoberta de um formalismo quimiométrico capaz de discriminar 38 indivíduos com câncer de próstata (CaP) e 23 sem a doença. Além disso, dos 38 indivíduos com CaP, 27, que seguiram sendo acompanhados foram divididos em três novos estudos metabonômicos preliminares: (1) estadiamento, composto por quinze pacientes com o índice de Gleason ≤ 2 e doze pacientes com Gleason > 2; (2) risco de recorrência, composto por doze pacientes com categoria T patológica, pT, < pT3 e margem cirúrgica negativa e quinze pacientes com classificação $\geq$ pT3 ou margem cirúrgica positiva; e (3) treze pacientes com remissão do câncer após a prostatectomia e catorze com recidiva bioquímica. Após essa busca exaustiva, para o diagnóstico de CaP, o formalismo com melhor desempenho foi o GA-LDA, que obteve sensibilidade e especificidade e exatidão, na validação externa, de 92%, 83% e 88%, respectivamente. A etapa de estadiamento, o modelo SFM-ETC-DT obteve precisão, sensibilidade e especificidade iguais a 88,9%, 75% e 100%, respectivamente. Para o risco de recidiva bioquímica, a partir da classificação pT e da margem cirúrgica, o modelo SFM-ETC-kSVM possuiu, na etapa de validação, precisão de 77,8%, sensibilidade de 75% e especificidade de 80%. Para a recidiva bioquímica, o modelo SFM-ETC-LDA obteve as figuras de mérito precisão, sensibilidade e especificidade, 80%, 87,5% e 60%, respectivamente. Para os seis conjuntos de dados avaliados, os modelos de classificação com regularização e treinados após seleção de variáveis e sobreamostragem obtiveram desempenho superior na generalização das predições do conjunto de validação, demonstrando a capacidade dessas técnicas na redução do sobreajuste. Mostrando também que a escolha apropriada de processamento de dados pode gerar modelos robustos, mesmo em dados pequenos e com grande dimensionalidade. Além disso, a técnica metabolômica baseada em RMN de ^{1}H se mostra uma alternativa interessante na redução de problemas durante o diagnóstico de CaP, podendo ser utilizado como uma ferramenta não invasiva para o auxílio médico no acompanhamento da doença.

Palavras-chave: Metabolômica; Quimiometria; Neoplasia; Soro; Espectroscopia; Pré-processamento de dados

ABSTRACT

One of the challenges faced by metabonomic studies, using nuclear magnetic resonance (NMR) spectral data obtained from human biofluids, is the small number of samples compared to the number of generated variables. In addition to this, health data can also be imbalanced, leading to overfitting in chemometric models. Therefore, widely used machine learning techniques, such as feature selection, oversampling, and regularization, can be employed to mitigate these effects. In this study, aiming to demonstrate the capability of these techniques to address overfitting issues in two previously published sets of metabonomic data based on ^{1}H NMR, seven feature selection techniques and five classification methods after resampling were exhaustively tested, resulting in 43 possible combinations. Furthermore, for a novel dataset, the same exhaustive process was carried out to discover a chemometric formalism capable of discriminating between 38 individuals with prostate cancer (PCa) and 23 without the disease. Additionally, out of the 38 individuals with PCa, 27 who continued to be monitored were divided into three new preliminary metabonomic studies: (1) staging, consisting of fifteen patients with Gleason score ≤ 2 and twelve patients with Gleason score > 2; (2) risk of recurrence, composed of twelve patients with pathological category T, pT, < pT3, and negative surgical margin, and fifteen patients with classification $\geq$ pT3 or positive surgical margin; and (3) thirteen patients in cancer remission after prostatectomy and fourteen with biochemical recurrence. After this exhaustive search, for PCa diagnosis, the formalism with the best performance was GA-LDA, achieving sensitivity, specificity, and accuracy of 92%, 83%, and 88%, respectively, in external validation. For staging, the SFM-ETC-DT model achieved precision, sensitivity, and specificity of 88.9%, 75%, and 100%, respectively. Regarding the risk of biochemical recurrence, based on pT classification and surgical margin, the SFM-ETC-kSVM model demonstrated precision of 77.8%, sensitivity of 75%, and specificity of 80% in the validation stage. For biochemical recurrence, the SFM-ETC-LDA model obtained precision, sensitivity, and specificity figures of merit at 80%, 87.5%, and 60%, respectively. For the six evaluated datasets, regularization-based classification models trained after feature selection and oversampling demonstrated superior performance in generalizing predictions to the validation set, highlighting the effectiveness of these techniques in reducing overfitting. This emphasizes that appropriate data processing choices can yield robust models even with small and high-dimensional datasets. Furthermore, the metabonomic technique based on ^{1}H NMR proves to be an interesting alternative in addressing issues during PCa diagnosis, serving as a non-invasive tool for medical support in disease monitoring.

Keywords: Metabolomics; Chemometrics; Neoplasm; Serum; Spectroscopy; Data preprocessing

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, com o desenvolvimento da computação, cresceram as possibilidades de aplicação de métodos estatísticos que obtiveram êxito na análise da grande massa de dados gerados pelas operações diárias em diversas áreas do conhecimento. Aliado a dados espectrométricos, esses métodos estatísticos promoveram o nascimento da Quimiometria (Ferreira, 2015; Pinto, 2017a).

O desenvolvimento da espectroscopia de ressonância magnética nuclear (RMN), associado ao desenvolvimento dos métodos quimiométricos e aprendizagem de máquinas (em inglês: *machine learning*) lançou um novo olhar sobre a análise de biofluidos, possibilitando reconhecimento de padrões e classificação de perfis metabólicos para separação de grupos de estudo, por exemplo, para o diagnóstico precoce de doenças e o monitoramento dos pacientes, sem a necessidade expressa de quantificação de um determinado biomarcador (Nicholson e Lindon, 2008). Isso pode ser útil no desenvolvimento de novos métodos de diagnóstico de doenças, pois o organismo dos pacientes responde à patologia resultando em alterações no metaboloma, devido à homeostase (Delafiori *et al.*, 2021; Jaurila *et al.*, 2020). Estudar o perfil desse metaboloma torna, muitas vezes, possível a discriminação entre os grupos, isso é interessante principalmente quando se trata de doenças carentes em relação a identificação de novos biomarcadores, que proporcionem um bom rastreio. A área que se preocupa com o estudo do perfil do metaboloma à resposta de um estímulo ou alteração genética é chamado de metabonômica (Antcliffe e Gordon, 2016; Bjerrum, 2015; Nicholson e Lindon, 2008; Zhang *et al.*, 2019).

Porém, dados atrelados a saúde humana, como os dados metabonômicos, podem ter problemas relativos ao tamanho do espaço amostral. Isso é causado, em geral, por questões como a dependência da incidência da doença e do tipo de amostragem. Em contraponto, o espectro de ressonância magnética nuclear pode resultar num grande número de variáveis, que depende da resolução requerida pelo operador. Esses pequenos conjuntos de dados, com grande dimensionalidade, podem acarretar modelagens matemáticas com muita variância, reduzindo assim a generalização para amostras desconhecidas (Ko, Choi e Ahn, 2021; Lin *et al.*, 2018; Vabalas *et al.*, 2019).

Um outro problema muito comum entre os dados médicos é a ocorrência de dados desbalanceados, ou seja, dados em que uma das classes possui um maior número de exemplos que outra. Esse episódio pode acarretar modelos enviesados em torno do grupo majoritário,

com maior número de amostras, tornando o modelo potencialmente predisposto a sobreajuste (Beinecke e Heider, 2021).

No estado da arte da aprendizagem de máquina já existem diversas ferramentas que podem mitigar os problemas acarretados por conjunto de dados pequenos e desbalanceados (Ko, Choi e Ahn, 2021; Lin, Lin e Li, 2023; Maray *et al.*, 2023). Técnicas de redução de dimensionalidade, podem reduzir o impacto do aumento da variância em estimadores, retirando variáveis que só contribuem para o ruído na modelagem (Jia *et al.*, 2022). Já as técnicas de reamostragem são uma alternativa para conjuntos de dados desbalanceados, criando amostras sintéticas no grupo minoritário ou removendo amostras do grupo majoritário (Rodrigues, Luna e Pinto, 2023).

As ferramentas metabonômicas tem se mostrado úteis no diagnóstico, estadiamento, entre outras classificações acerca de diversos problemas, dentre eles doenças oncológicas (Chen *et al.*, 2016; Griffin, 2020; Neto *et al.*, 2020; Vandergrift *et al.*, 2018).

O câncer de próstata (CaP) é um dos mais mortais em todo o mundo. De acordo com a GLOBOCAN 2020, cerca de 7% das mortes por câncer em homens no mundo são causadas por CaP (Sung *et al.*, 2021). Globalmente, o nível de antígeno prostático específico (PSA, do inglês: *prostate-specific antigen*), glicoproteína normalmente expressa pelo tecido da próstata, no soro sanguíneo é utilizado como método de rastreio do CaP (Pérez-Rambla *et al.*, 2017).

Nas últimas décadas, há uma diminuição na mortalidade dos casos de CaP, muito devido a intensificação do rastreio da doença e melhora nos métodos terapêuticos. Porém, a utilização desse método atual de rastreio a partir dos níveis séricos de PSA, de forma isolada, aumenta a possibilidade de sobrediagnóstico e sobretratamento, principalmente a partir da cirurgia de prostatectomia radical, de casos que seriam mais indolentes, que podem causar efeitos colaterais como incontinência urinária e impotência (Carlsson e Vickers, 2020; Toth *et al.*, 2019). Além disso o PSA não possui correlação com o grau do câncer. O que acarreta a necessidade urgente de novos métodos mais específicos e menos invasivos para a detecção e estadiamento do CaP (Stabile *et al.*, 2020; Vandergrift *et al.*, 2018).

Sendo assim, o seguinte trabalho atua em duas frentes. A princípio demonstrar que a utilização de técnicas de seleção de variáveis e reamostragem são capazes de reduzir os problemas de sobreajuste em conjuntos de dados pequenos e/ou desbalanceados. E utilizar dessas técnicas para avaliar se há a possibilidade de discriminação estatística entre as classes de pacientes (a) portadores de câncer de próstata (CaP) e indivíduos saudáveis (b) indivíduos com menor e maior grau do CaP, (c) com menor e maior risco de recidiva bioquímica da CaP pós prostatectomia radical e (d) pacientes que apresentaram recidiva bioquímica e que não

apresentaram a recidiva, a partir de métodos quimiométricos, utilizando dados espectrais de RMN de ^{1}H obtidos a partir de amostras de soro sanguíneo.

2 OBJETIVOS

2.1 Objetivo Geral

Investigar (1) o impacto da redução de dimensionalidade, regularização e de um método de sobreamostragem no aumento da generalização da predição de modelos de aprendizagem de máquina e (2) discriminar, a partir da estratégia metabonômica baseada em espectroscopia de Ressonância Magnética Nuclear de ^{1}H, pacientes acometidos por câncer de próstata de indivíduos saudáveis, a partir da construção de um modelo quimiométrico de classificação. Além disso, em um estudo preliminar criar modelos metabonômicos baseados em espectroscopia de Ressonância Magnética Nuclear de ^{1}H capazes de: (3) estadiar pacientes acometidos de câncer de próstata, (4) avaliar o risco e (5) prognosticar a recidiva bioquímica de câncer de próstata.

2.2 Objetivos específicos

Para a realização dos objetivos gerais, primeiramente, fez-se necessária a (1) obtenção dos espectros de Ressonância Magnética Nuclear de ^{1}H do soro dos pacientes advindos da clínica de urologia saudáveis e com câncer de próstata. (2) A matriz de dados criada a partir desses espectros e mais dois conjuntos de dados já publicados na literatura foram carregados como entrada em um laço em linguagem *Python* onde foram testadas, de forma exaustiva, combinações de seleções de variáveis e modelos de classificação otimizados, após a reamostragem. Com o conjunto de dados inédito, (3) foi escolhido o melhor modelo de classificação para análise metabonômica e, a partir das variáveis de interesse, (4) foi realizada a análise dos metabólitos mais importantes para a discriminação das classes. Além disso, (5) foi criado um conjunto de dados, apenas com pacientes portadores de câncer de próstata, com três diferentes vetores de resposta associados ao (a) estadiamento da doença, (b) avaliação do risco e (c) prognóstico da recidiva bioquímica do câncer de próstata. Para esses últimos casos, o conjunto de dados (6) foi pré-processado e (7) criados modelos quimiométricos de classificação para cada um dos vetores de resposta.

3 FUNDAMENTAÇÃO TEÓRICA

Considerando que o estudo aqui apresentado tem um caráter multidisciplinar, esse capítulo será dividido em cada uma das áreas de conhecimento aplicadas para a elaboração do trabalho. Isso será feito de forma concisa, para que os leitores de todas as áreas possam se familiarizar com cada etapa. Esse trecho se inicia falando sobre o que é a metabonômica (3.1); a espectroscopia de ressonância magnética nuclear (RMN) (3.2), que é o método espectroscópico pelo qual foram obtidos os dados de estudo; e, por fim, quimiometria e os principais métodos utilizados no processamento de dados para tomadas de decisão nesse trabalho (3.3); e os dados importantes sobre o câncer de próstata (CaP) e seu diagnóstico, em que se baseia a discussão final desse trabalho (3.4).

3.1 METABONÔMICA

As ciências ômicas surgiram com os avanços do sequenciamento genético e o nascimento da sua primogênita, a genômica. Como a bioquímica celular em sistemas vivos é extremamente complexa, desenvolveram-se novas áreas que se encarregam do estudo de sistemas específicos, como: a proteômica, que estuda os níveis de proteínas ou a metabolômica e metabonômica que estudam o metaboloma. A metabonômica e a metabolômica são utilizadas para diagnósticos, desenvolvimento de fármacos, estudos forenses, entre outros. Dentre essas últimas ciências ômicas, destaca-se nesse trabalho a metabonômica (Lerche *et al.*, 2019; Nicholson e Lindon, 2008; Stagljar, 2016).

A metabonômica foi um termo proposto por Jeremy Nicholson, Elaine Holmes e John Lindon (1999), que há mais de duas décadas, permeia as publicações científicas. Baseado em trabalhos publicados principalmente no fim da década de 70 e início da década de 80, a metabonômica foi definida como: “a medida quantitativa da resposta metabólica multiparamétrica dinâmica de sistemas vivos a estímulos fisiopatológicos ou modificação genética”. (DANIELS, WILLIAMS e WRIGHT, 1976; Griffin, 2020; Nicholson, Buckingham e Sadler, 1983; Nicholson, Lindon e Holmes, 1999).

Outro importante ponto, que vai além da própria definição de metabonômica, é fazer sua distinção ao termo metabolômica. Apesar de serem utilizadas muitas vezes como sinônimos, há uma diferença filosófica primordial, a metabonômica se preocupa, principalmente, em traçar um perfil metabólico global, o que diferencia do procedimento de

quantificação e caracterização da filosofia metabolômica (Griffin, 2020; Nicholson e Lindon, 2008).

A variedade de metabólitos encontrados em um biofluido, o metaboloma, é influenciado por fatores genéticos, que por sua vez, podem ser afetados por motivos fenotípicos como alimentação, uso de drogas, entre outros. Dessa forma, uma doença, por exemplo, acarretará uma mudança metabólica. Mesmo partindo de um sistema complexo e dinâmico, como o perfil metabólico de um ser humano, essa perturbação pode ser explorada pela metabonômica, e a partir de ferramentas estatísticas, pode se reconhecer padrões na alteração do equilíbrio metabólico, podendo criar uma espécie de "impressão digital" do metaboloma dos indivíduos acometidos por determinada doença, quando comparado com indivíduos saudáveis (Antcliffe e Gordon, 2016; Neto *et al.*, 2020).

A metabonômica pode ser aplicada com uma gama de ferramentas de análise, dentre elas podemos citar, principalmente, a espectroscopia de ressonância magnética nuclear (RMN) (Gómez-Cebrián *et al.*, 2019). Porém, além da espectroscopia de RMN, são encontrados trabalhos que utilizam a espectrometria de massas, a espectroscopia no infravermelho e Raman (Griffin, 2020). A técnica de RMN avançou significativamente desde a década de 1940, quando começou a ser difundida, trazendo uma maior quantidade de informação espectral útil acerca dos perfis metabólicos, tornando a RMN uma interessante aliada à filosofia metabonômica, principalmente pela facilidade na preparação das amostras, a alta reprodutibilidade e, em alguns casos, a grande velocidade da análise (Bjerrum, 2015). A relação entre os avanços da técnica analítica, que fornece os dados e o crescimento das técnicas de bioinformática, que se seguiram à crescente capacidade de armazenamento e processamento de dados por computadores em geral, além do desenvolvimento de métodos estatísticos robustos, foram cruciais para o crescimento da metabonômica (Griffin, 2020; Pinto, 2017b; Zhang *et al.*, 2018).

Sendo assim, a metabonômica é constituída das etapas de coleta de amostras de biofluidos, obtenção dos dados espectrais e seu pré-tratamento; a identificação dos metabólitos de interesse no espectro; a etapa estatística ou quimiométrica, etapa que será descrita em detalhes nos próximos itens; e, por fim, a tomada de decisão acerca dos resultados estatísticos obtidos (Bjerrum, 2015).

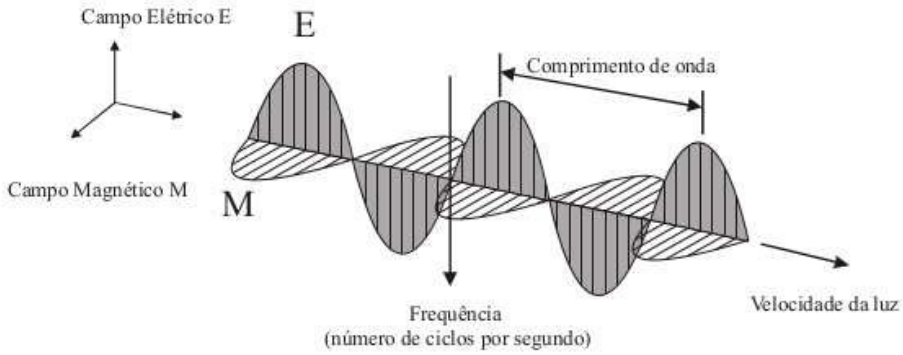
3.2 ESPECTROSCOPIA DE RESSONÂNCIA MAGNÉTICA NUCLEAR

A espectroscopia de ressonância magnética nuclear (RMN, do inglês: *Nuclear Magnetic Resonance*) tornou-se um método interdisciplinar poderoso. Pode se contar nove ganhadores do Prêmio Nobel desde a época em que Isador I. Rabi desenvolveu métodos de

ressonância para registrar as propriedades magnéticas de núcleos atômicos e recebeu o Prêmio Nobel de Física de 1944 (Ross *et al.*, 2007). Vários desenvolvimentos foram realizados no campo da RMN, como a incorporação da Transformada de Fourier e a utilização da técnica para obtenção de imagens corporais, com fins médico, nos anos 1980 (Ross *et al.*, 2007). Mas, antes de falarmos sobre o método espectroscópico, começaremos elucidando como o sinal espectroscópico é criado.

A espectroscopia de ressonância magnética nuclear é o estudo das transições energéticas associadas ao momento angular de núcleos atômicos de moléculas, quando submetidos a um campo externo, causadas pela absorção de radiação eletromagnética na região de radiofrequência (RF) (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016). A radiação é dita como eletromagnética pois, a radiação pode ser considerada como uma forma de onda constituída por duas componentes perpendiculares (Figura 1), um campo elétrico oscilante e um campo magnético oscilante.

Figura 1. Propagação da radiação eletromagnética com as indicações dos campos elétrico (E) e magnético (M).



Fonte: (Meneses *et al.*, 2012)

A radiação eletromagnética possui, além da característica de onda, características de partículas. Cada quanta da radiação é chamada de fóton, e cada fóton possui uma quantidade discreta de energia, E, que é diretamente proporcional a frequência de oscilação, ν, como descrito na Equação 1 (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016). A radiação na faixa de RF é uma radiação de baixa energia, como indicado no Quadro 1.

$$E = h\upsilon \quad (1)$$

Onde h é a constante de Planck.

Quadro 1. O espectro eletromagnético.

Radiação	Comprimento de onda (nm)	Frequência (10^6 Hz)	Energia (kJ · mol^{-1})
Raios gama	10^{-1} a 10^{-3}	10^{12} a 10^{14}	$1{,}2 \cdot 10^6$ a $1{,}2 \cdot 10^8$
Raios-X	10 a 10^{-1}	10^{10} a 10^{12}	$1{,}2 \cdot 10^4$ a $1{,}2 \cdot 10^6$
Ultravioleta	380 a 10	$8 \cdot 10^8$ a 10^{10}	$3{,}2 \cdot 10^2$ a $1{,}2 \cdot 10^4$
Luz visível	780 a 380	$4 \cdot 10^8$ a $8 \cdot 10^8$	$1{,}6 \cdot 10^2$ a $3{,}2 \cdot 10^2$
Infravermelho	10^5 a 780	10^6 a $4 \cdot 10^8$	0,4 a $1{,}6 \cdot 10^2$
Micro-ondas	10^5 a 10^7	10^4 a 10^6	$4 \cdot 10^{-3}$ a 0,4
Radiofrequência	10^{11} a 10^7	1 a 10^4	$4 \cdot 10^{-7}$ a $4 \cdot 10^{-3}$

Fonte: Adaptado de (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016)

A interação de cada tipo de radiação com a matéria é característica. No caso da ressonância magnética nuclear, a absorção de radiação na região de RF por átomos ativos submetidos em um campo magnético, fornece informações sobre ambiente molecular ao qual esse átomo está inserido (Solomons e Fryhle, 2005).

Esse efeito de ressonância magnética nuclear só ocorre devido ao núcleo possuir um momento angular, $\vec{I}$, mais conhecido como *spin* (KEELER, 2002). *Spin*, do inglês, significa "giro", a rigor, esse nome surgiu devido a ideia de que os elétrons possuíam um movimento de rotação, porém esse movimento não faz sentido quando atribuído ao núcleo. Portanto, o *spin* nuclear, para esse caso, será colocado como um quarto número quântico característico de cada núcleo atômico.

Além do *spin*, também é característico a existência de um momento magnético $\vec{\mu}$, dado pela Equação 2. Quando o núcleo do átomo possui um número par de prótons e nêutrons, o momento angular, é nulo, logo o momento magnético também se anula. Caso contrário, se o momento angular for diferente de zero, o núcleo possui um $\vec{\mu}$ diferente de zero e proporcional ao $\vec{I}$ (Bisht *et al.*, 2021).

$$\vec{\mu} = \hbar \vec{I} \gamma_X \qquad (2)$$

Onde ħ é a constante de Planck dividido por 2π e γ_X é a constante magnetogírica, que é uma característica própria de cada núcleo, que é definido como a razão entre o momento magnético e o momento angular.

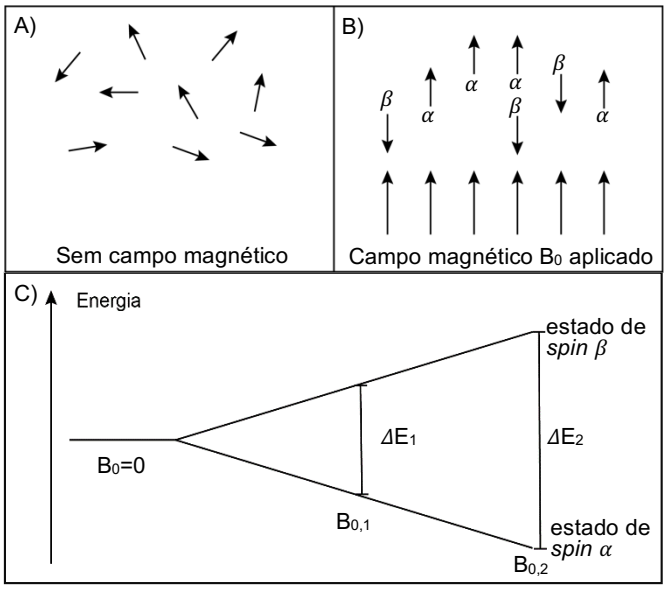
O *spin* nuclear pode adotar $(2I + 1)$ orientações quantizadas. Para um núcleo de *spin* 1/2, que é o caso do ^{1}H, ^{13}C, ^{19}F, ^{31}P, alguns dos núcleos mais importantes, principalmente para o estudo de moléculas orgânicas, temos os dois estados $+I$ e $-I$, de mesma energia[1] (Kupče *et al.*, 2021). Quando o núcleo está sob um campo magnético externo (B_0) (Figura 2), há a quebra

[1] Nesse trecho, *I* é o número quântico de *spin* nuclear.

da degenerescência[2] dos níveis de *spin* nuclear e nesse caso há agora uma diferença de energia, ΔE, entre esses dois estados, que é proporcional à B_0, como descrito na Equação 3. Sendo assim, cada núcleo atua como um pequeno imã e pode possuir o momento magnético alinhado ao campo B_0, então, temos um estado de *spin* de menor energia; ou o momento magnético pode estar alinhado contrariamente ao campo B_0, tendo assim um estado de *spin* de maior energia (Figura 2C) (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016; Ross *et al.*, 2007; Solomons e Fryhle, 2005).

$$\Delta E = \hbar \gamma_X B_0 \quad (3)$$

Figura 2. A) Na ausência de campo magnético externo os momentos magnéticos se orientam aleatoriamente. B) Com a inserção de um campo magnético externo B_0 os momentos magnéticos se orientam na direção paralela ao campo. C) Diagrama de energia entre dois estados de spin em função do campo magnético externo B_0.



Legenda: O estado de *spin* α, refere-se aos núcleos com momento magnético alinhados "a favor" o campo B_0. O estado de *spin* β, refere-se aos núcleos com momento magnético alinhados "contra" o campo B_0. Para um campo $B_{0,2} > B_{0,1}$, observa-se o crescimento da diferença de energia ($\Delta E_2 > \Delta E_1$)

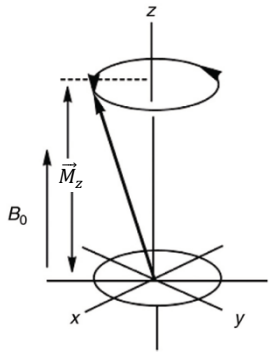
Fonte: Adaptado de (Solomons e Fryhle, 2005).

O vetor de magnetização dos núcleos se orienta no sentido do campo magnético, ou contrariamente a ele, conforme a Figura 2B, porém ele faz um movimento circular chamado de precessão em uma determinada frequência (Figura 3), conhecida como frequência de Larmor (Equação 4), ω_0. Se for aplicado um pulso de RF na direção perpendicular ao campo B_0, que coincida com essa frequência de precessão, por isso o nome "ressonância", pode ocorrer de núcleos mudarem do estado de *spin* de menor energia para o de maior energia. Nesse caso, temos que a energia do fóton E e o ΔE entre os níveis de *spin* devem ser equivalentes, então,

[2] Entende-se degenerescência como a existência, para um sistema quântico de estados diferentes, correspondendo a um mesmo nível de energia. A quebra da degenerescência, seria separar os estados, antes degenerados, em dois níveis energéticos distintos.

pode se igualar as Equações 1 e 3 e eliminando o valor da constante de Planck da equação, chega-se à equação fundamental da ressonância magnética nuclear (Equação 5) (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016).

Figura 3. Precessão do momento magnético no eixo do campo B_0 aplicado. A magnitude do vetor $\vec{M}_z$ corresponde ao excesso de núcleos de menor energia, de acordo com a distribuição de Boltzmann[3].



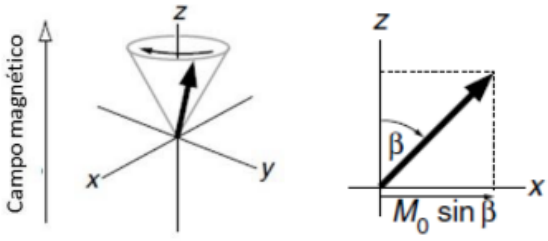
Fonte: (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016).

$$\omega_0 = \gamma_X B_0 \quad (4)$$

$$\upsilon = \frac{\gamma_X B_0}{2\pi} \quad (5)$$

Até agora, foi descrita principalmente a magnetização da amostra pelo campo B_0. Num sistema sem a presença do pulso de RF, tem-se um vetor magnetização resultante, $\vec{M}_0$, apontando para a direção *z*, que foi convencionada como a direção de B_0. Ao aplicar o pulso de RF, que também possui uma componente magnética, se tem a aplicação de um segundo campo magnético, B_1, ortogonal a *z*, capaz de afastar um pouco o vetor magnetização da direção do eixo *z*. Como a precessão continua, tem-se agora uma precessão mais afastada de *z*, com um ângulo constante, como mostra a Figura 4 (Ross *et al.*, 2007).

Figura 4. Do lado esquerdo: a precessão de Larmor após a aplicação de um campo magnético no plano xy. Do lado direito: a componente x do vetor magnetização M_0, com um ângulo constante β.



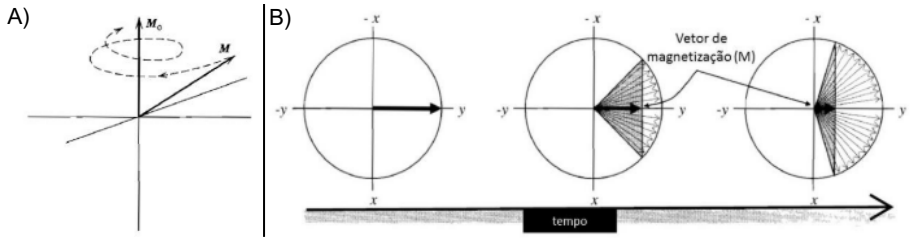
Fonte: Autoria Própria.

[3] A distribuição de Boltzmann segue a equação: $\ln(N_\alpha/N_\beta) = -\Delta E/(k_B T)$, onde N_α e N_β são o tamanho de duas populações em dois estados de energia distintos, ΔE é a diferença de energia entre os dois estados, k_B é a constante de Boltzmann e T é a temperatura absoluta.

O sistema de detecção é localizado no plano *xy*, por isso, os pulsos de RF são aplicados ortogonalmente ao eixo *z*. O vetor magnetização $\vec{M}_0$ pode ser decomposto em duas componentes $\vec{M}_{xy}$ e $\vec{M}_z$. Se não houver nenhuma radiação de RF o vetor em $\vec{M}_z$ é máximo, enquanto o $\vec{M}_{xy}$ é mínimo. Se um pulso de RF capaz de deslocar o vetor magnetização a 90°, for aplicado, por exemplo, tem-se, durante a aplicação do pulso, o $\vec{M}_{xy}$ no valor máximo e $\vec{M}_z$ é mínimo (Barros, 2017).

Uma vez que o pulso de RF é cessado, o sistema tende a retornar para o estado inicial, esse processo é chamado de relaxação. Essa relaxação pode ocorrer a partir do mecanismo *spin*-rede (T_1), que é a transferência de energia para o conjunto de moléculas vizinhas – a rede, também conhecida como relaxação longitudinal, que é o tempo necessário para que o vetor magnetização retorne à condição inicial no eixo *z* (Figura 5A). Há também o mecanismo *spin-spin* (T_2), também conhecido como relaxação transversal, que envolve a transferência de energia para os núcleos vizinhos, que pode provocar alargamento dos picos de espectros de RMN e perda de sinal, e corresponde ao tempo necessário para que a componente do vetor magnetização no plano *xy* chegue a zero (Figura 5B) (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016; Ross *et al.*, 2007).

Figura 5. A) Relaxação longitudinal, T_1, do vetor M para M_0. B) Relaxação transversal, T_2, do vetor magnetização no decorrer do tempo.



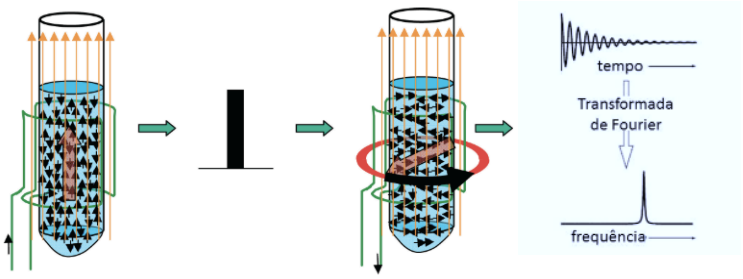
Fonte: Adaptado de KEELER (2002).

Para moléculas menores, em geral, $T_1 \approx T_2$. Já para macromoléculas $T_1 \gg T_2$, adicionalmente a isso, sinais de moléculas grandes decaem muito mais rápido que moléculas menores, pois o T_2 é inversamente proporcional ao raio hidrodinâmico da molécula e da viscosidade do meio. Essa característica pode ser utilizada na supressão de sinais de proteínas em estudos metabonômicos, como será visto mais a frente (Ross *et al.*, 2007).

A amostra é colocada dentro de uma bobina detetora (Figura 6). De acordo com a lei da indutividade, a precessão da magnetização induz uma diferença de potencial (ddp) modulada com a frequência. A amplitude dessa ddp é diretamente proporcional ao vetor $\vec{M}_{xy}$ e com o número de *spins* nessa frequência. A mudança na impedância das bobinas do oscilador causada pelas relaxações é medida pelo detector como um sinal na forma de um padrão de batimento

decadente no domínio do tempo, conhecido como decaimento livre da indução (FID, do inglês: *Free Induction Decay*) (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016; Ross *et al.*, 2007).

Figura 6. Detecção de um sinal de RMN e a transformada de Fourier

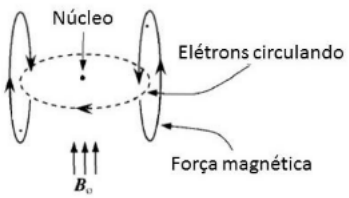


Nota: (esquerda) uma amostra com muitos *spins* (>10^{12}) é colocada dentro de um forte campo magnético externo B_0 (laranja) orientado ao longo da direção z. A amostra está contida em uma bobina de detecção (mostrada em verde). No equilíbrio, cerca de um *spin* em cada 10.000 contribui para uma magnetização macroscópica M_0 mostrada em vermelho transparente. Após a curta aplicação de um campo B_1 de alta frequência (pulso de 90°: barra preta), a magnetização é alinhada ao longo do eixo x do campo B_1 rotativo e começa a precessão em torno de B_0. Há uma tensão induzida na bobina de detecção (azul escuro) ao cessar o pulso de RF inicia-se a relaxação e o decaimento livre de indução (verde).
Fonte: Adaptado de ROSS (2007).

O sinal do FID, como mostrado na Figura 6, é um decaimento na ddp em função do tempo. Ao aplicar a transformada de Fourier, pode se obter sinais no domínio das frequências de ressonância dos núcleos em questão (Figura 6).

Como apresentado até agora, todos os sinais de RMN do mesmo núcleo atômico deveriam apresentar apenas um sinal, devido a frequência de Larmor ser característico do núcleo atômico e proporcional ao B_0 (Equação 4) (Atta-ur-Rahman, Choudhary e Atia-tul-Wahab, 2016). Por exemplo, para um núcleo de ^{1}H, num campo magnético B_0 de aproximadamente 9 T (tesla), a frequência de RF que possui quantidade de energia necessária para a transição do estado de menor energia para o de maior energia é de aproximadamente 400 MHz. Porém, o núcleo atômico não "sente" totalmente a influência de B_0. Há uma blindagem provocada pela densidade eletrônica do ambiente químico em que está o núcleo atômico. Essa blindagem decorre de um campo magnético induzido pelos elétrons e esse campo induzido opõe-se ao campo magnético externo (Figura 7) (Solomons e Fryhle, 2005).

Figura 7. Blindagem do núcleo por elétrons circulando ao seu redor.



Fonte: (Gouveia, 2017).

Essa blindagem depende da densidade eletrônica ao redor do núcleo. Essa última é influenciada por efeitos indutivos e anisotropia magnética. Sendo assim, átomos de maior eletronegatividade são capazes de reduzir a densidade eletrônica, "desblindando" o núcleo atômico, assim como correntes de elétrons π, também afetam a densidade eletrônica sobre o núcleo (Solomons e Fryhle, 2005). Esses efeitos de blindagem e desblindagem, que dependem do ambiente químico da molécula em que se encontra o núcleo e da sua vizinhança é o que caracteriza os sinais encontrados num espectro de RMN. A utilização da frequência de RMN efetiva, υ_{ef}, poderia ser bastante inconveniente, pois a mudança de frequência promovida pelo ambiente químico é da ordem de Hz (hertz), enquanto B_0 é da ordem de MHz (megahertz). Com isso, os espectros são mostrados usando uma escala adimensional chamada de deslocamento químico, δ, em partes por milhão (ppm), usando um sinal de referência, conforme a Equação 6.

$$\delta = 10^6 \times \frac{\upsilon_{ef} - \upsilon_{ref}}{\upsilon_{ref}} \quad (6)$$

Onde υ_{ref} é a frequência de ressonância para um composto de referência, como o tetrametilsilano (TMS), para solventes orgânicos, ou o ácido 3-trimetilsililpropanóico (TSP), para soluções aquosas (Ross *et al.*, 2007; Solomons e Fryhle, 2005). O espectro de RMN é plotado em função de δ, ao invés da frequência efetiva, o que torna o gráfico mais informativo e torna-o comparativo independentemente do campo B_0, pois tanto a frequência efetiva, quanto a frequência de referência variam com B_0.

É importante destacar o fato de que os núcleos, em uma espécie química, estão em ambientes eletrônicos diferentes em relação aos outros, apresentando frequências de ressonância distintas. Além disso, outros diversos efeitos de interação entre os campos magnéticos dos núcleos presentes na molécula e a possibilidade de análise quantitativa a partir da integral dos picos do espectro, que promove uma medida relativa das quantidades dos núcleos de hidrogênio, além de várias outros experimentos, tornam a espectroscopia de RMN uma ferramenta extremamente útil em diversas áreas do conhecimento científico (Barros, 2017; Gouveia, 2017).

O principal uso da RMN é para a elucidação estrutural de compostos. Para isso, várias sequências de pulsos 1D e 2D são empregadas, como: COSY (*COrrelated SpectroscopY*), HSQC (*Heteronuclear Single Quantum Correlation*), NOESY (*Nuclear Overhauser Effect SpectroscopY*), entre outros. Para a metabonômica, além das medidas de concentrações relativas, é importante também a identificação dos metabólitos. Esse último pode ser realizado, utilizando as sequências indicadas anteriormente. No entanto, há bancos de dados, como o

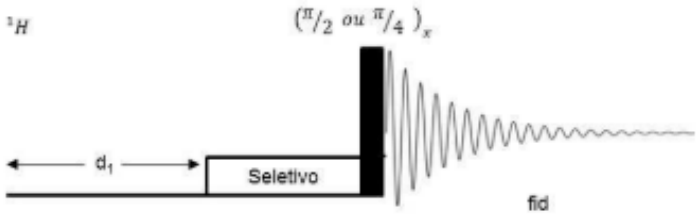
HMDB (do inglês: *Human Metabolome DataBase*), que podem ser utilizados com essa finalidade.

3.2.1 Sequência de pulsos de Pré-Saturação e CPMG

Como já citado anteriormente, a análise metabonômica utiliza como matriz biofluidos que são analisados, geralmente, usando espectroscopia de RMN. Porém, esse tipo de amostra possui alto teor de água como solvente, que dificulta a observação de compostos minoritários e dos metabólitos presentes nessa matriz.

A sequência de pulsos com pré-saturação do sinal do solvente (PRESAT), que pode ser utilizada para minimizar esse problema, consiste em dois pulsos de RF (Figura 8). O primeiro deles é um pulso de baixa intensidade na frequência de ressonância característica do solvente, chamado de pulso de saturação, que tem duração o suficiente igualar as duas populações de estado de spin, anulando o sinal de RMN do solvente ao longo da direção do campo gerado pela radiação de RF, do segundo pulso e é largamente utilizada para amostras de biofluidos, devido ao alto teor de água (Liu e Mao, 1999; Silva, 2017).

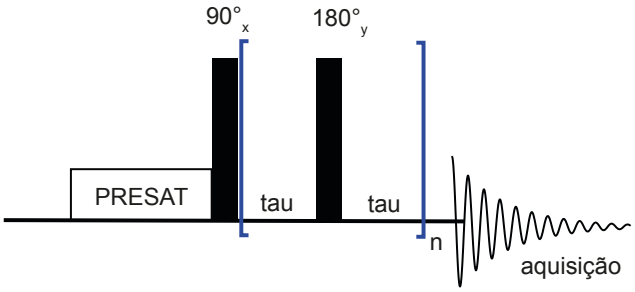
Figura 8. Sequência de pulsos de pré-saturação. Em destaque o pulso seletivo de pré-saturação.



Fonte: (Silva, 2017)

Porém, amostras como o soro de sangue, possuem um alto teor de macromoléculas, o que provoca sinais largos e pouco resolvidos na região de menores deslocamentos químicos (δ 0.80 – 1.30 ppm), que pode mascarar sinais de metabólitos menores. Esse tipo de inconveniente pode ser evitado explorando o fato de que macromoléculas possuem baixo valor de T_2. Métodos que reduzem esse efeito causado por macromoléculas em espectros de RMN são chamados de filtro de T_2 (Lucas *et al.*, 2005; Silva, 2017). Na redução desse outro problema, no presente trabalho foi escolhida a sequência de pulsos Carr-Purcell-Meiboom-Gill (CPMG), como um filtro de T_2, precedido por um pulso seletivo para a supressão do sinal da água (Figura 9)

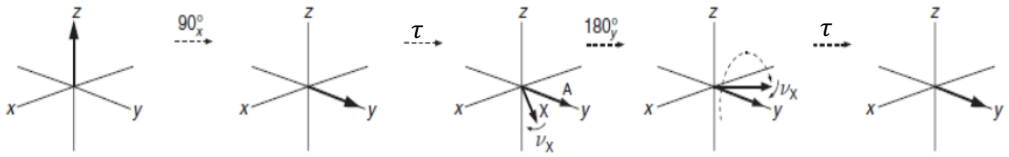
Figura 9. Sequência de pulsos PRESAT/CPMG



Fonte: Autoria própria.

A sequência CPMG consiste em um pulso de 90°, aplicado no eixo x, e sucessivos pulsos de 180° entre intervalos de tempo τ (tau), no eixo y, como mostrado na Figura 9. O vetor magnetização, antes da aplicação da sequência de pulsos, está alinhado no sentido do eixo $+z$. Quando aplicado um pulso de radiofrequência de 90º, ao longo do eixo x, o vetor magnetização tem sua orientação alterada e passa a ser analisado considerando as componentes no eixo z e no plano xy. Após um certo tempo, τ, aplica-se um pulso de 180º, no eixo y, com isso, inverte-se o sentido de precessão do vetor magnetização, produzindo um eco de spin. Em seguida, o equilíbrio é restabelecido e o vetor tende a voltar para a condição inicial (Figura 10) (Silva, 2017).

Figura 10. Vetores de magnetização e a sequência de pulsos CPMG.



Fonte: (Silva, 2017)

Moléculas grandes, em geral, possuem tempos de relaxação T_2 menores, voltando mais rapidamente ao seu estado inicial. O filtro de T_2 consiste em ajustar o valor de τ de tal forma que seja maior que o T_2 das espécies que se pretende eliminar e menor que o T_2 das espécies de interesse, deixando o espectro mais resolvido (Bjerrum, 2015; Ross *et al.*, 2007).

O RMN é um método analítico largamente utilizado, porém, para biofluidos, há diversas informações sobrepostas, que podem ser complexas numa análise inicial (Bisht *et al.*, 2021). Por conta disso, são utilizadas ferramentas estatísticas (quimiométricas) para extrair um maior número de informações úteis a partir desses espectros (Neto *et al.*, 2022).

3.3 QUIMIOMETRIA

O crescimento da química analítica confunde-se com a própria ciência química, que se desenvolveu desde a época em que se utilizava alfarrobas para cálculo de quilates do ouro, passando pela Idade Média e a alquimia, a criação dos métodos tradicionais por via úmida (titrimetria, gravimetria, precipitação etc.), o nascimento dos métodos espectroscópicos com Kirchoff e Bunsen, entre tantos outros avanços, que culminaram na grande tempestade de dados que hoje é encontrada nos métodos analíticos modernos (Adams e Adriaens, 2020; Karayannis e Efstathiou, 2012).

Muitos desses métodos citados acima, fazem parte da área da Química Analítica e do seu desenvolvimento. Durante as primeiras décadas do século XX já era conhecida a preocupação de químicos analíticos com questões estatísticas, como erros, acurácia e precisão, que com o passar dos anos foram incorporadas ao fluxo de trabalho. No fim da década de 1940, já se dava início a utilização do método de regressão por mínimos quadrados e análise de variância (ANOVA). Enquanto isso, durante as décadas de 1920 e 1930, métodos mais sofisticados de reconhecimento de padrões, como a análise de componentes principais e outros métodos de discriminação eram desenvolvidos e utilizados principalmente nas áreas de ciências sociais e psicologia (Brereton *et al.*, 2017).

Acompanhando o crescimento das capacidades de processamento de computadores e a evolução dos equipamentos analíticos, em 1969, Jurs, Kowalski, Isenhour e Reilly publicaram uma sequência de artigos sobre aprendizagem de máquinas em problemas químicos. Era a vanguarda da união entre métodos de cálculos computacionais e a química (Ferreira, 2015). Svante Wold, em 1971, utilizou pela primeira vez a palavra *Kemometri*, associando métodos estatísticos avançados, matemática computacional e problemas de natureza química. Já em 1975, surgiu pela primeira vez a palavra *Chemometrics*. E em 10 de junho de 1974, juntamente com Bruce Kowalski, Wold criou a *International Chemometric Society* (Kowalski, Brown e Vandeginste, 1987; Kowalski, 1975).

Mesmo com vários problemas em relação a padronização, que ocorrem ainda hoje, a quimiometria se difundiu e no final do ano de 1980, o próprio Bruce Kowalski ministrou um curso no Instituto de Química da Universidade Estadual de Campinas, Unicamp. Após isso, foi iniciada a formação de diversos grupos da área, tanto no Brasil, quanto na América Latina (Ferreira, 2015).

A partir dos trabalhos de Kowalski, Massart e outros, em seu livro intitulado *Handbook of Chemometrics and Qualimetrics: Part A*, desenvolveu a definição de quimiometria como:

"Quimiometria é a disciplina que usa a matemática, a estatística e a lógica para: a) planejar ou otimizar procedimentos experimentais; b) extrair o máximo da informação química relevante, através da análise dos dados; e c) obter conhecimento sobre sistemas químicos." (Ferreira, 2015; Lavine, 2018)

A extração dessas informações continua movendo cientistas ao redor do mundo, desenvolvendo novos métodos matemáticos. A Quimiometria pode ser dividida em três grandes áreas: a calibração multivariada, reconhecimento de padrões e planejamento de experimentos. Além disso, existem diversos métodos "acessórios", que otimizam as metodologias supracitadas, como o pré-processamento e pré-tratamento de dados, entre outros (Lavine e Workman, 2013). Nesse trabalho, o foco principal será na abordagem dos métodos de reconhecimento de padrões e pré-processamento de dados.

Os métodos de reconhecimento de padrões podem ser divididos em duas classes, os **não-supervisionados** e os **supervisionados**. Por não-supervisionados, entende-se como os métodos que não utilizam informação *a priori* das classes as quais pertencem as amostras no momento da utilização do algoritmo matemático (Ferreira, 2015). A análise de componentes principais (PCA, do inglês: *Principal Component Analysis*) ou os métodos aglomerativos, que podem ser utilizados para criar grupos de classificações indicativas de filmes em plataformas *on-line* como o *Netflix*, são exemplos de métodos não-supervisionados (Raschka e Mirjalili, 2019).

Já os métodos de classificação, ou métodos de reconhecimento supervisionados, utilizam um número de classes pré-estabelecidas durante a construção do modelo matemático, essa etapa é chamada de treinamento. Uma vez construído esse modelo, o método deve ser capaz de classificar novas amostras – amostras de teste – como pertencentes ou não às classes propostas previamente ao treinamento (Gao *et al.*, 2019; Raschka e Mirjalili, 2019).

Uma importante questão que envolve os algoritmos de aprendizagem de máquinas é o equilíbrio entre o viés e a variância explicada pelo modelo. Por viés do modelo, entende-se como o erro de previsão do modelo, que causa a perda das relações entre as variáveis e o vetor alvo. Já no caso contrário, um modelo com alta variância explicada, considera o ruído no treinamento, criando um modelo excessivamente complexo, causando **sobreajuste**, que leva a não generalização do modelo quando utilizadas amostras de teste (Raschka e Mirjalili, 2019).

Alguns desses métodos podem ser ditos como paramétricos ou não-paramétricos, esse conceito surge da existência ou não de assunção de que a distribuição dos resultados é conhecida *a priori* (Duda, Hart e Stork, 2000). Alguns métodos como a Regressão Logística

(LR, do inglês: *Logistic Regression*) e a Análise Discriminante Linear (LDA, do inglês: *Linear Discriminant Analysis*) são métodos paramétricos (Nanga *et al.*, 2021; Raschka e Mirjalili, 2019). Já os métodos como os k-Vizinhos Mais Próximos (KNN, do inglês: *k-Nearest Neighbors*), os métodos de classificação e regressão baseados em Árvores de Decisão são exemplos comuns de métodos não-paramétricos, que não fazem inferência em relação a nenhum parâmetro relacionado à distribuição dos dados, como média, covariância etc. (Chong, Wishart e Xia, 2019; Ferreira, 2015; Reenen, Van *et al.*, 2017).

Na continuação desse item, vamos abordar os fundamentos das etapas e os modelos utilizados na construção desse trabalho, iniciando pelos métodos de pré-tratamento e pré-processamento de dados e os métodos de reconhecimento de padrão utilizados.

3.3.1 Pré-tratamento e Pré-processamento de Dados

Uma vez obtidos e organizados os dados de maneira adequada, esses podem necessitar de um processamento prévio a análise multivariada. Essa etapa pode reduzir influências indesejadas ao modelo, que comprometeriam a análise dos resultados e levariam a tomadas de decisão equivocadas.

Diversos métodos matemáticos computacionais foram desenvolvidos para pré-processar matrizes de dados analíticos, e é um desafio a qualquer utilizador escolher o caminho que torna mais amigável a matriz de dados ao método quimiométrico empregado. Vale a pena ressaltar alguns desses métodos como, os métodos de **suavização, correção de linha de base**, **normalização**, **autoescalamento**, **alinhamento**, **seleção de variáveis**, **reamostragem,** etc. (Bjerrum, 2015, p. 123; Guha *et al.*, 2021; Pandey e Janghel, 2019). Vamos citar alguns desses métodos a seguir.

Os sinais analíticos são constituídos a partir da contribuição de dois efeitos. O sinal verdadeiro, devido a concentração do analito, e o ruído, que é uma componente aleatória. Visando reduzir o efeito do ruído, os métodos de **suavização** podem ser realizados pelo próprio equipamento em que foi realizada a análise ou posteriormente pelo analista após a aquisição dos dados. Nos métodos espectroscópicos, a **suavização** tem como objetivo aumentar a razão sinal/ruído (S/R) do sinal analítico em questão. Entende-se por S/R como o quociente entre o sinal analítico e a raiz quadrada do erro quadrático médio – ou raiz da variância do sinal (Ferreira, 2015).

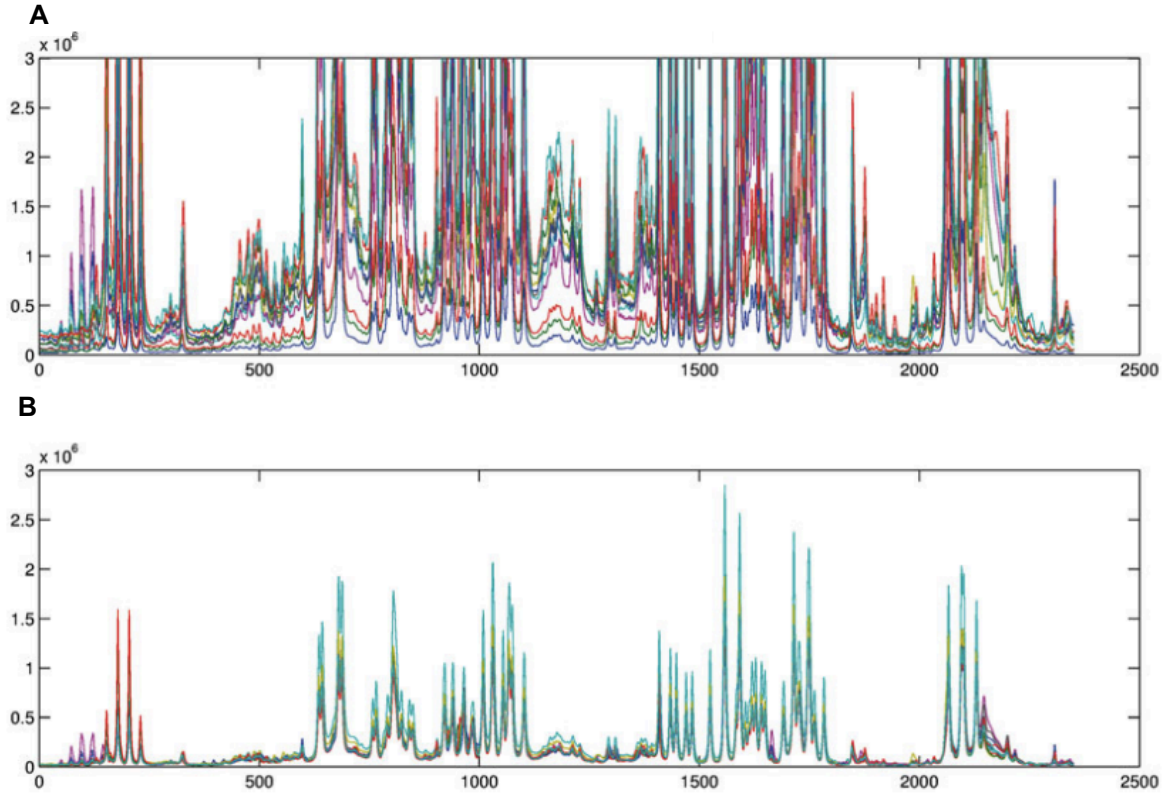
Além das contribuições aleatórias, informações sistemáticas também podem atrapalhar a construção de modelos quimiométricos. A distorção da linha de base, um

deslocamento em relação ao eixo das ordenadas em espectros, pode afetar tanto a construção de modelos estatísticos, quanto a quantificação dos analitos (Bjerrum, 2015, p. 127). Essa distorção pode ser corrigida a partir de métodos de derivada de **correção de linha de base** que acompanham alguns equipamentos analíticos, ou que podem ser realizadas posteriormente pelo analista, com ajuda de *softwares* de pré-processamento de dados (Emwas *et al.*, 2018).

Outra informação importante é o possível deslocamento no eixo das abcissas. Esse tipo de deslocamento pode ocorrer em espectros de RMN, por exemplo. Esse problema pode ser reduzido ou eliminado quimiometricamente por um método de alinhamento. Existem diversos desses métodos na literatura, para os dados de RMN é muito utilizado o método, em inglês, *binning* ou *bucketing*, que consiste em dividir o espectro em janelas (*bins*) de mesma largura e somar as intensidades, ou áreas sob o espectro. O *binning* pode contornar o problema de alinhamento, além disso pode também reduzir a contribuição dos ruídos (Ferreira, 2015).

Devido a possibilidade de ocorrência de uma grande diferença de concentração entre amostras de biofluidos (Figura 11), como é o caso desse trabalho, algumas amostras podem ter maior magnitude do sinal que em outros na mesma região, provocando dificuldades à construção dos modelos estatísticos. Por conta disso, uma etapa importante do pré-processamento de dados é a **normalização** desses dados, que deve ser realizado após qualquer etapa de **limpeza de dados**. Por limpeza de dados entende-se a retirada de regiões específicas do espectro ou da matriz de dados que não possuam valores significativos, ou seja, sejam ruídos, ou que possuam informação não importante para a construção do modelo quimiométrico – como o sinal da água em espectros de RMN (Emwas *et al.*, 2018; Ferreira, 2015).

Figura 11. Exemplo de normalização de dados quimiométricos. A) sem a normalização e B) após a normalização.



Fonte: (Bjerrum, 2015)

Retornando ao tema de normalização, destacaremos o método conhecido pelo termo em inglês *Standard Normal Variate* (SNV). No SNV, cada valor em uma linha de dados da matriz, x_{ij}, é subtraída pela média de cada linha, $\bar{x}_i$, e dividida pelo respectivo desvio padrão dessa linha da matriz, s_i, como indicam as Equações 7, 8 e 9 abaixo (Ferreira, 2015).

$$\bar{x}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij} \tag{7}$$

$$s_i = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2} \tag{8}$$

$$x_{ij,snv} = \frac{1}{s_i}(x_{ij} - \bar{x}_i) \tag{9}$$

Onde x_{ij} são os dados da matriz, J é o número total de variáveis e j é a identificação da j-ésima variável, $x_{ij,snv}$ são os valores da matriz após a normalização.

Outro problema comum em análise de dados é a diferença entre as dimensões dos dados em cada variável, j, o que pode afetar significativamente a análise, elevando demais a importância de uma variável em relação a outra. Para tentar minimizar ou eliminar esse empecilho, será destacado o uso do método de **autoescalamento** (Emwas *et al.*, 2018).

O autoescalamento é um processo em que os dados centrados na média são divididos pelo valor do desvio padrão dos dados calculados em cada variável, s_j, como mostrado nas equações 13 e 13 abaixo (Chong, Wishart e Xia, 2019).

$$s_j = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ij} - \bar{x}_j)^2} \tag{13}$$

$$x_{ij,sc} = \frac{1}{s_j}(x_{ij} - \bar{x}_j) \tag{14}$$

Onde, $x_{ij,sc}$ é o dado após o autoescalamento. Diferentemente da centragem na média, o autoescalamento torna os dados adimensionais, retirando sua variação com respeito a unidade original dos dados (Ferreira, 2015).

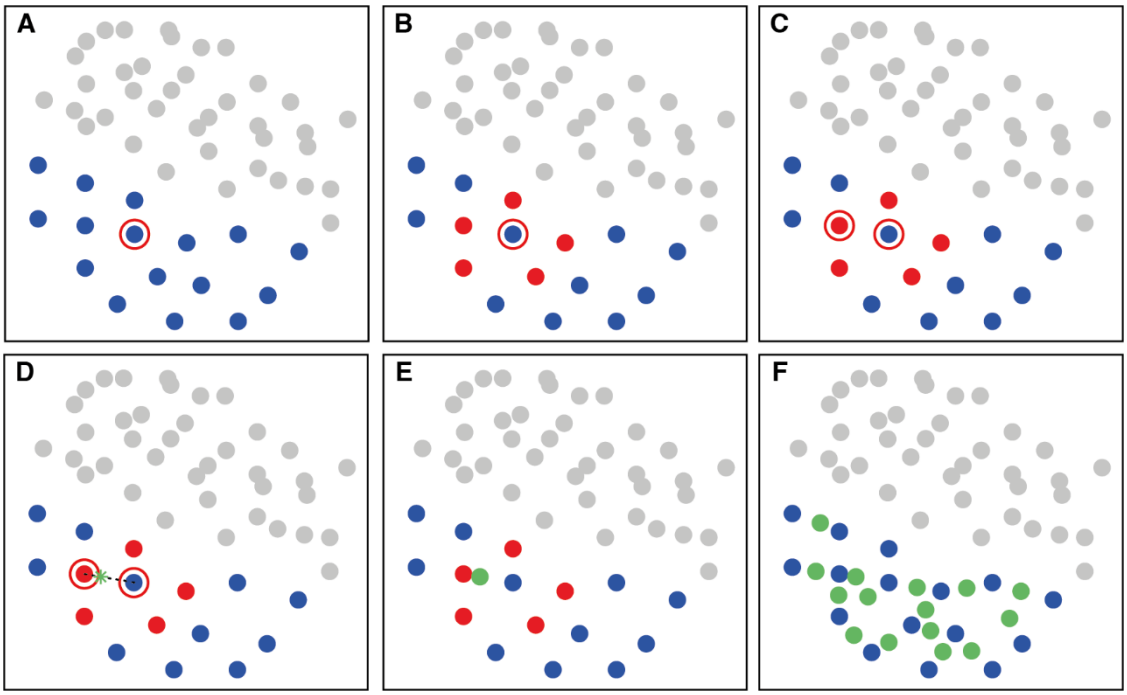
No pré-processamento de dados há outros processos aplicados para resolução de questões específicas como dados que possuem desequilíbrio entre o número de amostras em cada classe. Um dos métodos utilizados para solucionar esse tipo de problema é a técnica conhecida em inglês como *Synthetic Minority Over-sampling TEchnique* (SMOTE), que gera amostras sintéticas na classe com menor número de amostras, a fim de equilibrar as classes do conjunto de dados, sem duplicar amostras originais (Pandey e Janghel, 2019).

Um conjunto de dados pode ser considerado desbalanceado ou desequilibrado, se houver um número de amostras menor em uma das classes (classe minoritária) que em outra classe (classe majoritária). Em aprendizado supervisionado, esse problema pode causar um viés em relação à classe majoritária, além disso, como a maioria dos algoritmos usam a exatidão como parâmetro de desempenho, as classes desbalanceadas podem reduzir o poder de avaliação da exatidão (Rodriguez-Torres, Carrasco-Ochoa e Martínez-Trinidad, 2019; Umer *et al.*, 2021). Por exemplo, um conjunto de dados de teste com 1100 amostras possui 1000 amostras do grupo positivo e 100 do grupo negativo. Se o algoritmo de classificação errar 50 amostras do grupo negativo, este terá aproximadamente 95% de exatidão, mesmo classificando erroneamente 50% das amostras do grupo negativo.

Para contornar esse tipo de problema, Chawla *et al*. (2002) propuseram um algoritmo estocástico que adiciona ao grupo minoritário amostras sintéticas localizadas em pontos aleatoriamente distribuídos em meio a distância euclidiana entre as amostras e os vizinhos mais próximos, do grupo a ser sobreamostrado. O processo realizado pelo algoritmo está apresentado na Figura 12, abaixo.

Figura 12. Funcionamento do algoritmo SMOTE. (A) Inicialmente o algoritmo seleciona uma amostra aleatória e em seguida, (B) computa seus k-vizinhos mais próximos. (C) Ainda de forma aleatória, o algoritmo escolhe um dos k-vizinhos mais próximos e (E) gera uma amostra sintética numa (D) distância

intermediária aleatória entre a amostra e o vizinho escolhido. (F) O processo é iterativo até alcançar uma condição estabelecida pelo operador.



Fonte: Autoria própria.

Como descrito no trabalho de Chawla e colaboradores (2002), a primeira etapa (Figura 12A) consiste em selecionar cada amostra do grupo minoritário de forma aleatória. Para cada uma das amostras selecionadas dessa classe, são calculados os k-vizinhos mais próximos, nesse exemplo, cinco, que se apresentam em vermelho na Figura 12B. Uma vez computadas as distâncias para cada um dos vizinhos mais próximos, é escolhido de forma aleatória um desses vizinhos (Figura 12C). A partir da distância a esse vizinho, é escolhido um valor, também de forma aleatória, entre 0 e 1 para multiplicar por essa distância, criando assim um ponto intermediário entre a amostra e seu vizinho (Figura 12D). Então, é adicionada uma amostra sintética nesse ponto escolhido (Figura 12E). O processo é então repetido até atingir o número desejado de amostras sintéticas a serem geradas (Figura 12F).

Apesar da importância do SMOTE na redução da variância do modelo e aumento da generalização, em alguns casos, essa técnica possui algumas limitações, como a incapacidade de considerar o ruído das amostras ou não considerar a fronteira entre as duas classes na geração de amostras, por conta disso, existem diversas outras técnicas de sobreamostragem, muitas delas são adaptações ao próprio SMOTE com o objetivo de sanar algumas dessas demandas (Beinecke e Heider, 2021; Rodriguez-Torres, Carrasco-Ochoa e Martínez-Trinidad, 2019).

3.3.2 Seleção de Variáveis

Além dos demais métodos de pré-processamento já citados, na resolução de questões de aprendizagem de máquina ou quimiometria, sejam elas de classificação ou regressão, uma importante etapa em grande parte dos casos é a seleção de variáveis. Essa etapa consiste em reduzir o espaço de variáveis, mantendo aquelas mais informativas em relação ao método de estimação proposto. A seleção de variáveis ajuda na melhora do desempenho na predição do método de classificação ou regressão, reduzindo então a possibilidade de sobreajuste, como também no tempo de processamento, ao diminuir o tamanho da dimensão da matriz de dados a ser computada (Guha *et al.*, 2021).

Os métodos de seleção de variáveis são divididos em três tipos, os **métodos de filtragem** (conhecidos em inglês como: *filter methods*), os métodos conhecidos em inglês como *wrapper* (traduziremos nesse texto como **métodos empacotados**), e os **métodos integrados** (conhecidos em inglês como: *embedded methods*). Essa última não será abordada no presente texto. Para mais informações ler (Jia *et al.*, 2022).

Nos **métodos de filtro,** o novo subconjunto de dados é escolhido sem a utilização de um método aprendizagem. Nesse caso, é escolhido um limite a partir de uma determinada avaliação das variáveis. Alguns dos métodos utilizados para fazer essa avaliação dos atributos do conjunto de dados são, por exemplo, distâncias estatísticas (como a Euclidiana ou de Mahalanobis), correlação entre as variáveis (também conhecida como multicolinearidade), um teste estatístico (como o teste F da ANOVA), a variância das variáveis, etc. (Hira e Gillies, 2015; Jia *et al.*, 2022).

Já os **métodos empacotados** incorporam um método de aprendizagem ao processo de seleção de variáveis. Ele classifica a quantidade de informação embutida em cada variável ou subconjunto de variáveis para formar a nova matriz com dimensão reduzida, utilizando alguma métrica do algoritmo de aprendizagem (Jia *et al.*, 2022).

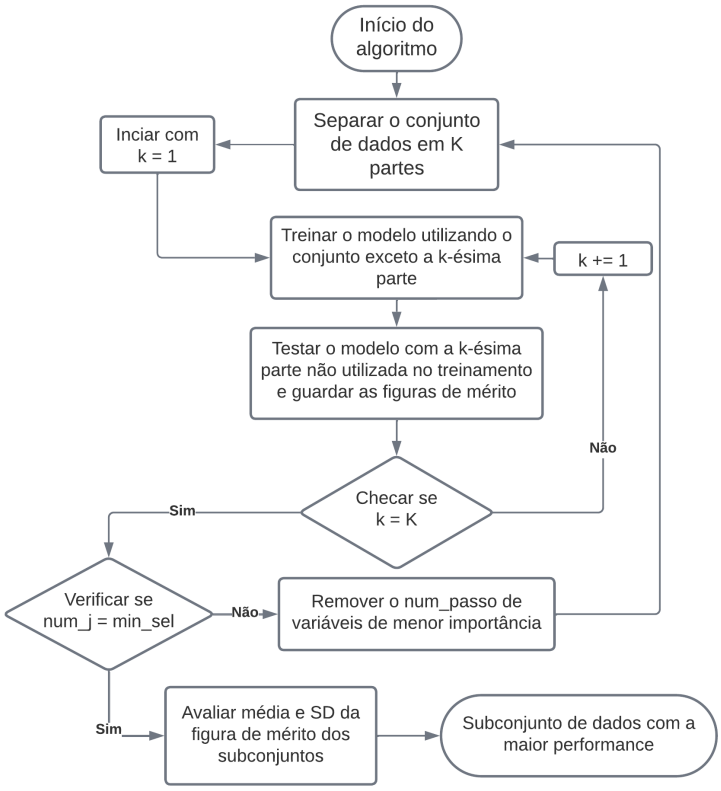
O método que solucionaria o problema de se obter o melhor subconjunto de dados para determinado estimador seria o método de busca exaustiva. Para um conjunto com J variáveis, o número possível de subconjuntos seria de $(2^J - 1)$. Esse número pode ser grande demais para ser computado, o que o torna um algoritmo difícil demais de praticar mesmo com os avanços em computação atuais (Galvão, Araújo, de e Soares, 2020). Porém, há outros métodos capazes de chegar a resultados satisfatórios, ainda que não analise todo o espaço de probabilidades.

Alguns exemplos, pertinentes a esse trabalho, dessa classe de metodologias de seleção de variáveis são: a eliminação recursiva de variáveis (RFE, do inglês: *Recursive Feature Elimination*), a seleção de variáveis sequencial (SFS, do inglês: *Sequential Feature Selection*)

e os métodos meta-heurísticos como o algoritmo genético (GA, do inglês: *Genetic Algorithm*) (Ba *et al.*, 2023; Khaire e Dhanalakshmi, 2022; Sadeghian *et al.*, 2023; Wei *et al.*, 2023).

O RFE é um método em que há eliminação sucessiva das variáveis menos importante ranqueadas a partir da importância da variável para o classificador ou regressor, até um número de atributos pré-determinado pelo operador (Kornyo *et al.*, 2023). Uma interessante implementação encontrada na biblioteca *scikit-learn* para linguagem *Python* é o RFECV() (Figura 13) (Pedregosa, F. *et al.*, 2011). Essa função é composta pelo RFE acrescido de uma camada de validação cruzada (CV, do inglês: *Cross Validation*). Nesse caso, a RFE com CV (RFECV) remove a etapa de escolha do hiperarâmetro número de atributos a escolher, deixando a CV determinar qual o melhor número de variáveis a ser escolhida pelo algoritmo (Awad e Fraihat, 2023).

Figura 13. Algoritmo RFECV implementado na biblioteca *scikit-learn*.



Legenda: ***num_j*** é o número de variáveis selecionadas até então e ***min_sel*** é o número mínimo de amostras a ser selecionado.

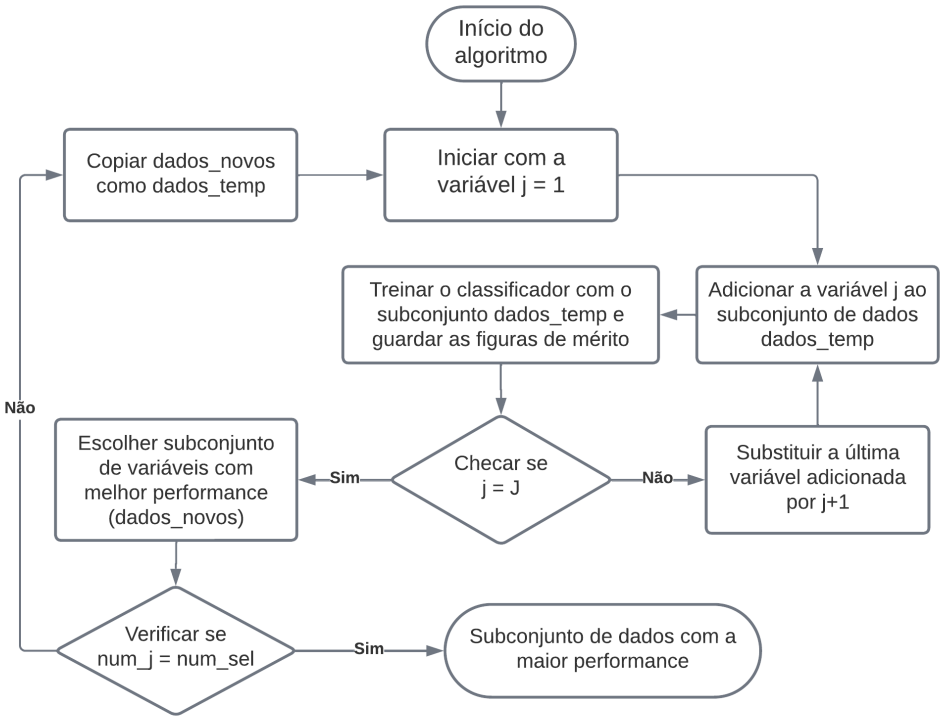
Fonte: Autoria própria.

A RFECV é um método mais custoso computacionalmente que o RFE padrão, porém, ele engloba o processo de escolha do melhor número de variáveis, reduzindo o processo de otimização do método (Awad e Fraihat, 2023).

O algoritmo SFS, é implementado na biblioteca *scikit-learn* com duas direções, avançado ou retrógado (Pedregosa, F. *et al.*, 2011). No método retrógrado, o algoritmo inicia com todas as J variáveis e computa o desempenho do algoritmo de aprendizagem. Então, o algoritmo remove uma das variáveis, e testa cada uma das possibilidades com J–1 variáveis. O subconjunto com dimensão $I \times J$–1, cujo estimador possuir a melhor performance é mantido. O algoritmo, então, remove mais uma das variáveis e refaz o processo anterior. Esse laço é realizado até o número de atributos pré-estabelecidos pelo operador ou o incremento na Figura de mérito ganha com a extração de uma nova variável seja menor que um limite também estabelecido pelo operador. Esse método demanda um grande esforço computacional, principalmente para dados com muitas variáveis (Jia *et al.*, 2022).

No método avançado (Figura 14), o algoritmo inicia com nenhuma variável e testa variável por variável até encontrar aquela cujo estimador possua o maior valor de figura de mérito. Escolhida a primeira variável, é refeito o processo para que seja adicionada outra variável. O laço se segue até encontrar o número de variáveis pré-estabelecidas ou o incremento na Figura de mérito ganha com a adição de uma nova variável seja menor que um limite (Jia *et al.*, 2022; Wei *et al.*, 2023).

Figura 14. Algoritmo do método SFS no método avançado.

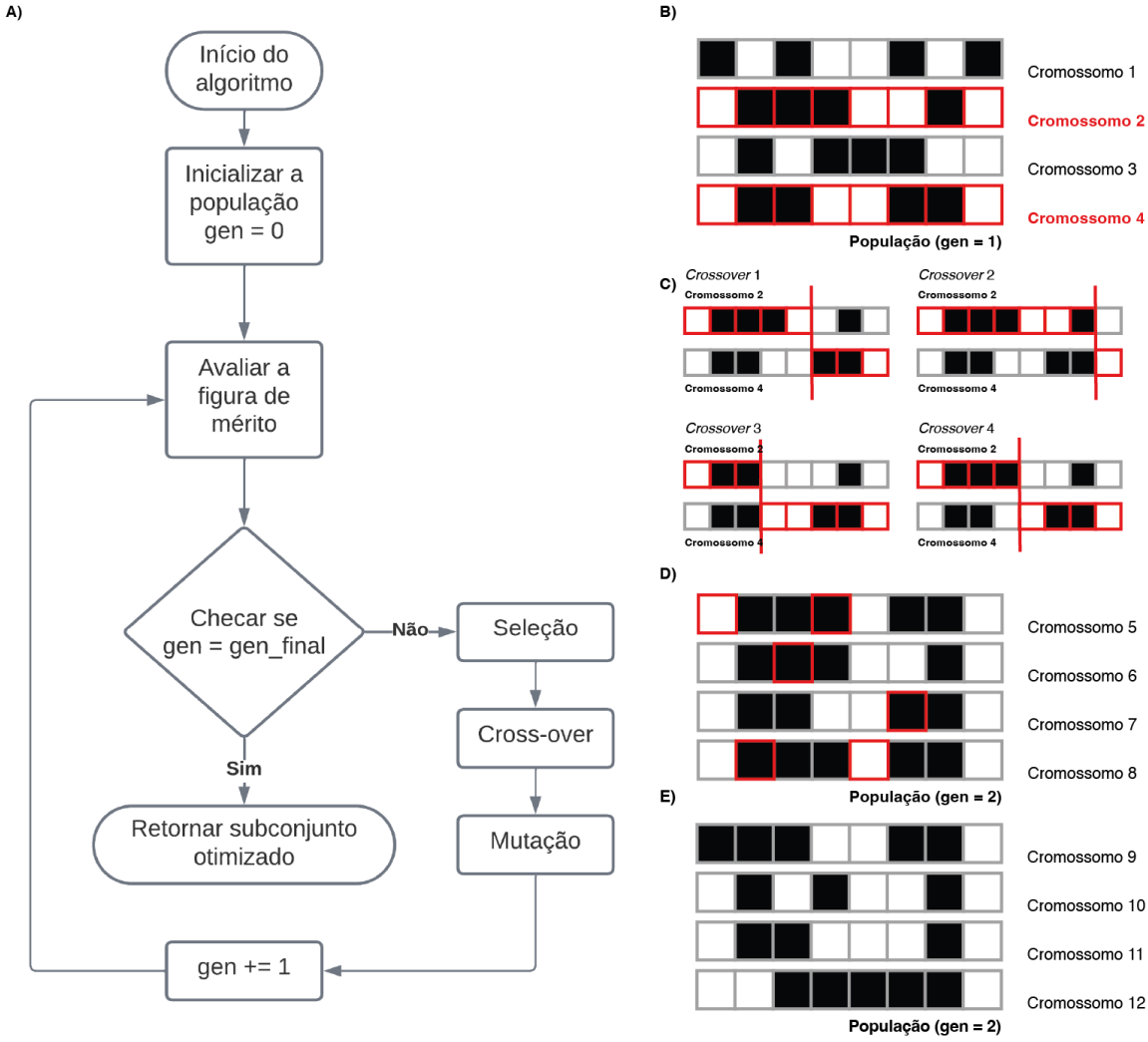


Legenda: ***J*** é o número total de variáveis do conjunto de dados, ***dados_temp*** é um subconjunto de dados temporários dentro do algoritmo, ***num_j*** é o número de variáveis selecionadas até então e ***num_sel*** é o número de variáveis que devem ser selecionadas ao término do algoritmo.
Fonte: Autoria própria.

O método SFS é um método chamado de “ganancioso”, uma vez que ele analisa um grande subespaço de probabilidades de combinações de variáveis, porém, ele é extremamente dependente do hiperparâmetro número de variáveis selecionados, o que pode custar uma redução em sua performance, se a escolha não for ideal (Jia *et al.*, 2022).

Em busca de resultados mais rápidos, apesar do grande número de informações atuais, algoritmos de otimização meta-heurísticos têm sido largamente utilizados por conta da sua simplicidade, ao utilizar paralelização e apenas uma função de ajuste. Nesse contexto, um dos mais conhecidos métodos meta-heurísticos de otimização é o GA. Esse algoritmo é um método baseado diretamente na seleção natural e é composto essencialmente por quatro funções (Figura 15) (Sadeghian *et al.*, 2023).

Figura 15. Algoritmo Genético, (A) fluxograma e suas principais funções. As funções são: (B) seleção, (C) cross-over e (D) a mutação, que (E) funcionam de forma iterativa.



Legenda: ***gen*** é a geração e ***gen_final*** é o valor escolhido de iterações totais.
Fonte: Autoria própria.

O GA inicializa criando uma população (matriz) de cromossomos aleatórios (Figura 15A). Cada cromossomo é um vetor composto por zeros e uns (genes), que são relativos à

aceitação ou não daquela variável no subconjunto (1 para verdadeiro, 0 para falso). (1) A primeira função é a função de ajuste, que é um algoritmo de aprendizagem que retorna uma determinada figura de mérito, como a exatidão num método de classificação. Uma vez inicializada a população, cada indivíduo, ou seja, cromossomo serve como uma máscara para selecionar um subconjunto de dados. Esse subconjunto é passado pela função de ajuste e retorna um valor que é computado. (2) A função de seleção seleciona aleatoriamente alguns cromossomos da população e faz um torneio entre eles. Os melhores desse torneio (Figura 15B) são agora os cromossomos pais da próxima geração (ou iteração). (3) No *crossover* (Figura 15C), partes dos cromossomos de cada par de pais é combinada em porções também aleatórias, formando novos indivíduos, que são a junção dos pais da geração anterior. (4) Os cromossomos da nova geração são mutados (Figuras 15D e 15E), ou seja, alguns genes são aleatoriamente modificados, para aumentar a variabilidade de resposta, diminuindo a probabilidade de o algoritmo ficar "preso" em um máximo local. Por fim, após o número de gerações selecionado, o melhor cromossomo, no último torneio, é escolhido para formar o subconjunto de dados final.

O GA mostra-se com diversas vantagens, como a boa paralelização, aplicabilidade, ser robusto em diversas aplicações e simples. Porém, apesar da mutação, um dos maiores problemas enfrentados pelo GA é a possibilidade de se "contentar" em um "topo de monte", que é apenas o máximo local, uma vez que a mutação não garante que a população vai levar a "descer o monte" e procurar um caminho alternativo na próxima geração (Guha *et al.*, 2021; Jia *et al.*, 2022).

Além dos métodos citados anteriormente, um outro método de seleção de variáveis utilizado é formar um subconjunto de dados a partir, diretamente, das j-variáveis mais importantes no treinamento de determinado estimador. Esse método é bastante empregado com classificadores como a Florestas Aleatórias e a Regressão Logística (Furmańczyk *et al.*, 2023; Gündoğdu, 2023; Speiser *et al.*, 2019; Vishraj, Gupta e Singh, 2023).

3.3.3 Análise de Componentes Principais

A Análise de componentes principais (PCA, do inglês, *Principal Component Analysis*) foi introduzida por Pearson, em 1901, porém o tratamento formal do método foi devido ao trabalho de Hotteling (1933), o que causou uma revolução no uso de métodos multivariados na área de psicologia (Ferreira, 2015).

A PCA é um dos métodos mais amplamente utilizado na quimiometria. É um método linear baseado em projeção e correlação. Por conta dessa simplicidade a PCA é muito utilizada como ferramenta na etapa de análise exploratória dos dados (EDA, do inglês: *Exploratory Data*

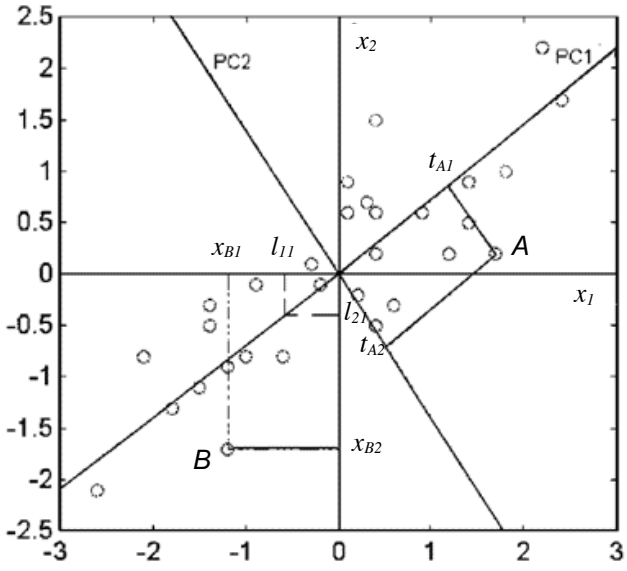
Analysis), pois é importante citar que, na análise multivariada incorre um problema de dificuldade de visualização dos dados, devido a grande quantidade de variáveis, para superar essa dificuldade, a PCA promove a redução da dimensionalidade dos dados, a partir da projeção desses em novas variáveis não-correlacionadas, chamadas de **Componentes Principais** (PC, do inglês: *Principal Component*), que são ortogonais entre si (Pinto, 2017b; Popovic *et al.*, 2019).

Tnovas variáveis, isso permite que a PCA reduza a dimensionalidade dos dados, descartando o ruído (Ferreira, 2015; Pinto, 2017b). Uma matriz **X** de dados é decomposta em *p* PCs a partir da Equação 15, abaixo:

$$\boldsymbol{X} = \boldsymbol{T_A L_A^T} + \boldsymbol{E} \tag{15}$$

Onde $\boldsymbol{T}$ é a matriz de escores e $\boldsymbol{L}$ é uma matriz ortonormal[4] de pesos (ou do inglês: *loadings*) e $\boldsymbol{E}$ é a matriz de erros, que possui as variâncias residuais. A matriz de escores expressa a relação entre as amostras e as PCs, já a matriz de pesos, a relação entre as variáveis da matriz de dados $\boldsymbol{X}$ e as PCs. Para ficar mais clara essa relação, observando a Figura 12 temos a construção de uma PCA com duas PCs (PC1 e PC2), a partir de uma matriz de dados bidimensional, com as duas variáveis x_1 e x_2. Nesse caso, verificamos que a amostra B, possui as coordenadas x_{B1} e x_{B2}, nas variáveis originais da matriz de dados $\boldsymbol{X}$.

Figura 16. Gráfico de um conjunto de dados bidimensionais (x_1, x_2), mostrando o eixo das componentes principais (PC1, PC2).



Fonte: FERREIRA et al. (1999).

[4] A matriz A é ortonormal quando $\mathbf{A^T A} = \mathbf{A A^T} = \mathbf{I}$, onde **I** é a matriz identidade – uma matriz em que a diagonal é composta por 1 e os demais valores são iguais a zero.

As PCs são construídas no sentido da máxima variação dos dados e no sentido de diminuir os resíduos. Sendo assim, a primeira PC (PC1) é a direção de máxima variação, como observado na Figura 12. Já a segunda componente principal (PC2) é a segunda com maior "explicação" da variação dos dados e ortogonal a PC1. Se seguir com um número maior de PCs, algumas delas terão pequenas variações, por conta disso há diversas formas de escolher o número de PCs adequadas ao modelo, como o método de validação cruzada ou simplesmente escolhendo um valor máximo de variação explicada por todas as PCs (95%, por exemplo). Retornando à Figura 9, uma vez que foram definidas as PCs, temos os valores de t_{A1} e t_{A2} como as projeções do ponto A, sobre os eixos das PC1 e PC2, nesse caso, são os escores da amostra A em cada uma das PCs. Já os valores l_{11} e l_{21} são os pesos das variáveis λ_1 e λ_2 na construção da PC1, que também pode ser avaliado como a influência da variável na construção da PC. (Antcliffe e Gordon, 2016; Ferreira, 2015).

Existem várias formas de calcular as matrizes $\boldsymbol{T}$ e $\boldsymbol{L}$, a mais comum delas é a decomposição em valores singulares (SVD, do inglês: *Singular Value Decomposition*). O SVD decompõe a matriz de dados inicial em três matrizes: $\boldsymbol{U}$, $\boldsymbol{S}$ e $\boldsymbol{V}$, como descrito na Equação 16 (Ferreira, 2015; Lay, 2011).

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{16}$$

Se compararmos as Equações 15 e 16, podemos observar que o produto $\boldsymbol{US}$ (I x J) é a matriz de escores $\boldsymbol{T}$ e a matriz $\boldsymbol{V}$ é a matriz de pesos $\boldsymbol{L}$. A matriz $\boldsymbol{SS}^T$ é uma matriz diagonal quadrada (I x I) e ainda segundo a Equação 17 mostra a relação entre essa matriz e a matriz diagonal de autovalores $\boldsymbol{\Lambda}$, que possuem os autovalores, λ_k. O valor λ_k corresponde à variância explicada pela k-ésima PC (Equação 19).

$$\mathbf{SS}^{\mathrm{T}} = \mathbf{\Lambda} \tag{17}$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \tag{18}$$

Uma outra forma de cálculo das matrizes da PCA é o algoritmo, cujo nome em inglês é *Non Linear Iterative Partial Least Squares* (NIPALS), que calcula os autovalores um a um e não todos de uma vez, como no SVD (Ferreira, 2015).

A PCA pode ser uma ferramenta de análise de amostras anômalas quando utilizada juntamente com uma elipse de confiança traçada a partir da estatística T2 de Hotelling. Mais detalhes são encontrados no capítulo 3 de (Ferreira, 2015).

Uma vez que foram citados os métodos de pré-processamentos e o método de EDA pertinentes ao presente trabalho, segue uma análise sobre os métodos de classificação aplicados.

3.3.4 Métodos de Classificação

Um método de classificação traz ao analista a possibilidade de inferir as possibilidades de uma amostra pertencerem, ou não, a um determinado conjunto, chamado aqui de grupo ou classe, que nada mais é que um conjunto de amostras que pertencem a uma condição pré-definida, por exemplo, portadores de uma patologia e amostras de controle ou um produto normatizado e outro, ou outros, fora da norma vigente etc.

Como citado anteriormente, os métodos de reconhecimento de padrão podem ser supervisionados ou não supervisionados. Nos próximos subitens serão discorridos sobre os métodos supervisionados de classificação que foram utilizados tanto na seleção de variáveis, ou como classificadores.

3.3.4.1 Regressão Logística – LR

A regressão logística (LR, do inglês: *Logistic Regression*) é um método de classificação de implementação simples e utilizado para classes linearmente separadas no hiperespaço dos dados. Por conta disso, é bastante utilizado em diversas aplicações e com algoritmos implementados em diversas bibliotecas e aplicativos (Raschka e Mirjalili, 2019).

A LR é derivada da função **logit** ou do logaritmo natural da chance (em inglês: *odds*) (Equação 19), que é o logaritmo da probabilidade de um evento positivo, por exemplo, uma amostra pertencer a uma determinada classe, ocorrer (Raschka e Mirjalili, 2019).

$$logit(p) = \ln(odds) = \ln\left(\frac{1-p}{p}\right) \tag{19}$$

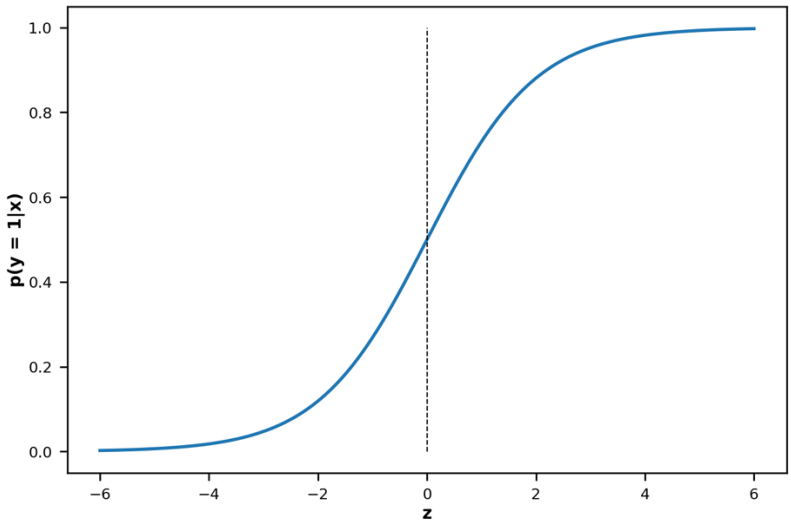
No gráfico da função logit, os valores da variável independente percorrem os valores 0 a 1, porém para fazer com que ela seja uma função que acesse a probabilidade de pertencer a determinada classe na variável dependente, é necessário tomar o inverso da função (Equações 20 e 21).

$$\hat{p}(\boldsymbol{x_i}) = \mathrm{p}(y = 1|x) = \mathrm{logit}^{-1}(z) = \frac{1}{1-e^{-z}} \tag{20}$$

$$z = w_0x_0 + w_1x_1 + \cdots + w_Jx_J = \boldsymbol{w}^T\boldsymbol{x} \tag{21}$$

Nessas equações, $\hat{p}(X_i)$ e $\mathrm{p}(y = 1|x)$ são a probabilidade de um dado vetor (amostra), $\boldsymbol{x_i}$, pertencer à classe 1, *w* representa os coeficientes lineares da função e *J* é o número total de variáveis. Apesar do nome regressão, ela é utilizada para delimitar duas classes. Porém, esse nome vem do fato de aplicar uma função sigmoidal a uma regressão linear múltipla. A função sigmoidal **logit^{-1}** possui o aspecto mostrado na Figura 17 (Daniel e Martin, 2023).

Figura 17. Forma da função sigmoidal que representa a probabilidade de um exemplo *x* pertencer à classe "*y* = 1*"*, numa regressão logística.



Fonte: Autoria própria.

O cálculo dos coeficientes da matriz de pesos $\boldsymbol{W}$ é calculada de forma iterativa a fim de reduzir uma função de custo, conhecida como função de verossimilhança, $L(\boldsymbol{w})$ (Vishraj, Gupta e Singh, 2023). Em geral, a versão do logaritmo natural da verossimilhança, é mais utilizada (Equação 22) (Raschka e Mirjalili, 2019).

$$\ln\left(L(\boldsymbol{w})\right) = C\sum_{i=1}^{I}\left(-y_i \ln\left(\hat{p}(\boldsymbol{x_i})\right) - (1-y_i)\ln\left(1-\hat{p}(\boldsymbol{x_i})\right)\right) + r(w) \tag{22}$$

Nessa equação, y_i é o valor da resposta da classe verdadeira para cada amostra. O *r(w)* é termo de regularização e C é o termo que determina a "força" dessa regularização. O *r(w)* serve para "penalizar" as variáveis com coeficientes, *w*, maiores, a fim de evitar um aumento da variância do modelo, reduzindo o risco de sobreajuste. Já o termo C denota o quanto a regularização será levada em consideração na função de custo. Valores menores de C indicam uma maior importância da regularização (Pedregosa, F. *et al.*, 2011). Dois dos mais frequentes termos de regularização encontrados nos pacotes computacionais mais comuns são as regularizações L1 e L2, encontradas nas equações 23 e 24, respectivamente (Raschka e Mirjalili, 2019).

$$r(w)_1 = \|\boldsymbol{w}\| \tag{23}$$

$$r(w)_2 = \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{24}$$

Esse artifício é muito utilizado para redução de sobreajuste de modelos e é encontrado também em outros modelos de classificação, como o SVM, que será citado mais a diante. Essas

regularizações possuem efeitos muito distintos e a escolha deles deve ser realizada de forma cautelosa. Para mais detalhes, ler (Daniel e Martin, 2023; Raschka e Mirjalili, 2019).

3.3.4.2 Análise Discriminante Linear – LDA

A LDA é um algoritmo supervisionado, mas que, assim como a PCA, promove a redução de dimensionalidade de um conjunto de dados a partir de uma função linear. A LDA foi proposta pela primeira vez por Fischer em 1936 com a proposta de classificar um problema binário (Nanga *et al.*, 2021). A redução da dimensionalidade promovida por esse algoritmo segue os passos citados no quadro abaixo (Quadro 2).

Quadro 2. Sequência de cálculos do algoritmo de redução de dimensionalidade LDA.

1. Normalizar o conjunto de dados.
2. Computar o vetor média para cada classe *k* ($\boldsymbol{m_k}$)

$$m_k = \frac{1}{n_k}\sum x_k \tag{25}$$

3. Construir a matriz de dispersão intraclasse, $\boldsymbol{S_w}$.

$$S_w = \sum_{k=1}^{K} S_k \tag{26}$$

$$S_k = \sum_{x \in D_k} (x - m_k)\,(x - m_k)^T \tag{27}$$

4. Construir a matriz de dispersão entre-classes, $\boldsymbol{S_b}$.

$$S_b = \sum_{k=1}^{K} N_i (m_k - m)\,(m_k - m)^T \tag{28}$$

5. Computar os autovetores e autovalores da matriz $S_w^{-1}S_b$.
6. Ordenar do maior para o menor os autovalores e seus respectivos autovetores.
7. Escolher os maiores autovalores. A partir desses autovalores construir a matriz W de transformação de dimensão J × K-1. Os autovalores são as colunas dessa matriz.
8. Projetar as amostras no novo subespaço a partir do produto entre a matriz X e W.

No Quadro 2, é possível identificar que após a normalização, o próximo passo para a realização do LDA é calcular a média dos dados de cada classe, $\boldsymbol{m_k}$, dividindo o somatório dos valores dessa classe $\boldsymbol{x_k}$ pelo número total de exemplos nessa classe n_k e K sendo o número de classes. A partir daí, serão calculadas as duas matrizes de dispersão: intraclasse, $\boldsymbol{S_w}$, e entre-classes, $\boldsymbol{S_b}$. A primeira também pode ser calculada a partir da matriz de covariância da classe, $\sum_k$, como descrito na equação 29. Já a matriz $\boldsymbol{S_b}$ é calculada como a diferença entre o vetor médio da classe, $\boldsymbol{m_k}$, e o vetor médio do conjunto de dados, $\boldsymbol{m}$.

$$S_w = \sum_{k=1}^{K} (n_k - 1)\Sigma_k \quad (29)$$

Assim como pode ser utilizado para redução de dimensionalidade, como descrito no Quadro 1, a LDA pode ser utilizado como método de classificação, acessando a probabilidade de predição do modelo, a partir do Teorema de Bayes (Equação 30) (Duda, Hart e Stork, 2000). Além da assunção anterior, é também condicionado que a distribuição condicional de classe P(x|y=k) é modelado como uma distribuição de Gauss e, em adição, a covariância de cada classe é igualada, $\sum_k = \sum$ e a função de discriminante linear (DF, do inglês: *Decision Function*) ou o "log[P(y=k|x)]" pode ser dado pela equação 31 ou pela equação 32. Mais detalhes em (Pedregosa, F. *et al.*, 2011; Tharwat *et al.*, 2017).

$$P(y = k|x) = \frac{P(x|y = k)P(y = k)}{P(x)} \quad (30)$$

$$DF = lnP(y = k|x) = -\frac{1}{2}(x - m_k)^T \Sigma^{-1}(x - m_k) + lnP(y = k) \quad (31)$$

$$DF = \boldsymbol{w_k^t x} + \boldsymbol{w_0} \quad (32)$$

O primeiro termo da equação 31 é a distância de Mahalanobis. Essa é a distância entre um determinado exemplo, $\boldsymbol{x}$, e a média do conjunto de dados da classe *k*, $\boldsymbol{m_k}$, levando em consideração a variância da variável. Já a equação 32, define a DF, onde o $\boldsymbol{w_k}$ é o vetor dos coeficientes lineares da função e o $\boldsymbol{w_0}$ é o intercepto da função (Duda, Hart e Stork, 2000).

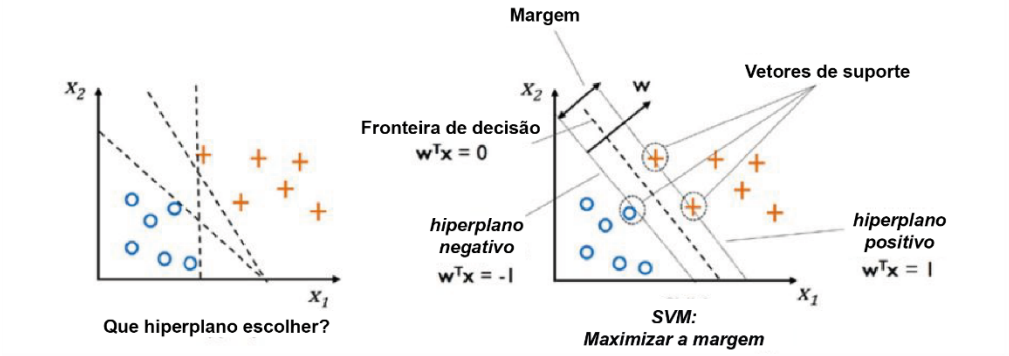
Apesar de ser muito empregado, a LDA possui algumas desvantagens. A primeira delas é que esse algoritmo, descrito acima, não pode ser utilizado quando a matriz de dados possui um número de variáveis maior que o número de exemplos, J > I. Além disso, o método é sensível a colinearidades no conjunto de dados e assume que as classes são linearmente separadas (Ferreira, 2015).

Na implementação no scikit-learn, o algoritmo LDA possui um método, conhecido em inglês como *shrinkage*, no português livre: **encolhimento**, como método de regularização, assim como os métodos L1 e L2 para LR. Esse método é indicado nos casos em que a matriz possui um número de exemplos próximo ao número de atributos ou variáveis (Pedregosa, F. *et al.*, 2011). Nesse caso, a covariância, utilizada pela LDA, é um estimador pobre (Ledoit e Wolf, 2004). O encolhimento utiliza uma combinação linear entre uma função estruturada, como proposto por Ledoit e Wolf (2004), e a matriz de covariância, desta feita, esse procedimento visa aumentar a generalização do classificador, evitando assim o superajuste.

3.3.4.3 Máquina de Vetores de Suporte – SVM

O método SVM é um algoritmo de aprendizagem largamente utilizado. Diferentemente do algoritmo LR, que tenta de forma iterativa reduzir as classificações erradas a cada passo da otimização da função de custo, no algoritmo SVM a otimização tem como objetivo aumentar a margem entre os grupos (Figura 18) (Duda, Hart e Stork, 2000; Raschka e Mirjalili, 2019).

Figura 18. As fronteiras, os vetores de suporte e a margem no algoritmo SVM.



Fonte: (Raschka e Mirjalili, 2019).

Na Figura 18, observa-se que a margem é a distância entre o hiperplano de separação entre as classes, a fronteira de decisão, e os exemplos que se encontram mais próximos a essa fronteira, os vetores de suporte, daí o nome do algoritmo (Raschka e Mirjalili, 2019). Os vetores de suporte delimitam dois hiperplanos paralelos à fronteira de decisão, estes estão descritos nas equações 33 e 34 e a generalização das duas descrito na equação 35.

$$\boldsymbol{w} \cdot \boldsymbol{x_{pos}} + \boldsymbol{b} = 1, se\ y = +1 \tag{33}$$

$$\boldsymbol{w} \cdot \boldsymbol{x_{neg}} + \boldsymbol{b} = -1, se\ y = -1 \tag{34}$$

$$y_k(\boldsymbol{w} \cdot \boldsymbol{x_k} + \boldsymbol{b}) - 1 \geq 0, para\ y_k = +1, -1 \tag{35}$$

Para essas equações, $\boldsymbol{x_{pos}}$ e $\boldsymbol{x_{neg}}$ são os exemplos que pertencem à classe positiva ou negativa, $\boldsymbol{w}$ é o vetor dos coeficientes dos hiperplanos, o índice *k*, indica a classe que pertence o vetor $\boldsymbol{x}$. A distância entre as margens pode ser calculada a partir da diferença entre as duas retas normalizadas pelo comprimento do vetor $\boldsymbol{w}$, $||\boldsymbol{w}||$, chegando na equação 36.

$$\frac{\boldsymbol{w} \cdot (\boldsymbol{x_{pos}} - \boldsymbol{x_{neg}})}{||\boldsymbol{w}||} = \frac{2}{||\boldsymbol{w}||} \tag{36}$$

Quanto maior a margem, maior o poder de generalização do modelo, uma vez que margens pequenas tendem a sobreajuste do modelo. Nesse caso, é interessante promover a maximização do segundo membro da equação 36. Essa maximização é equivalente a minimizar o termo "$||\boldsymbol{w}||^2/2$". Portanto chega-se ao problema de otimização que se encontra na equação 37.

$$minimizar\ \frac{\|\boldsymbol{w}\|^2}{2}, sujeito\ a\ y_k(\boldsymbol{w}\cdot\boldsymbol{x_k}+\boldsymbol{b})-1\geq 0, \qquad k=1,\dots K \tag{37}$$

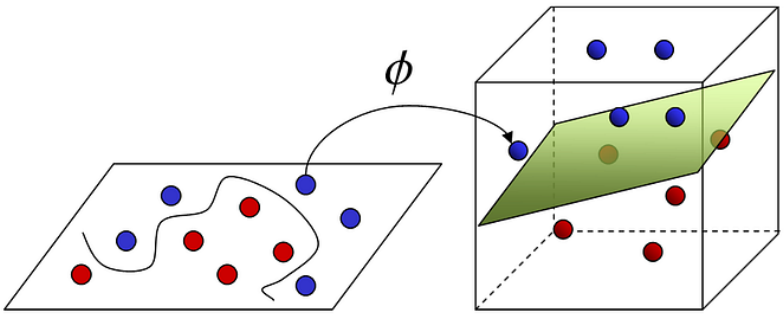
Para lidar com dados que não sejam totalmente não-linearmente separados, Cortes e Vapnik (1995) propuseram o método de **margem suave**, que acrescenta uma variável de folga, ξ. Além disso, em algumas aplicações também se adiciona um termo de regularização, C. Enquanto maior esse termo, maior a penalização do modelo por conta de erros de classificação, porém enquanto menor o valor, maior a margem, reduzindo assim os riscos de sobreajuste (Cortes e Vapnik, 1995; Duda, Hart e Stork, 2000; Pedregosa, F. *et al.*, 2011). Com a adição desses novos termos, as restrições para otimização passam a ser as encontradas na equação 38.

$$\begin{aligned} &minimizar\ \frac{\|\boldsymbol{w}\|^2}{2}+C\sum_{k=1}^{K}\xi_k\,, sujeito\ a\ y_k(\boldsymbol{w}\cdot\boldsymbol{x_k}+\boldsymbol{b})\geq 1-\xi_k, \\ &\xi_k\geq 0, k=1,\dots K \end{aligned} \tag{38}$$

A otimização a partir das restrições propostas na equação 38 não fazem parte do escopo deste trabalho e podem ser encontrados em (Vapnik, 2000).

O SVM, como um algoritmo baseado numa função linear, possui dificuldades em modelar matrizes de dados em que as classes não são linearmente separáveis. Para lidar com esse tipo de situação, foi criado o chamado truque do *kernel* (Raschka e Mirjalili, 2019). O conceito básico desse método é criar combinações não-lineares da função original para projetar em outro espaço de maior dimensão a partir de uma função de mapeamento, ϕ, onde os dados podem ser separados linearmente (Figura 19) (Yan e Zhu, 2022).

Figura 19. O truque do *kernel*. Os dados que não podem ser separados linearmente são transformados a partir de uma função ϕ, que projeta os dados em um espaço de maior dimensão que pode ser separado por um hiperplano.



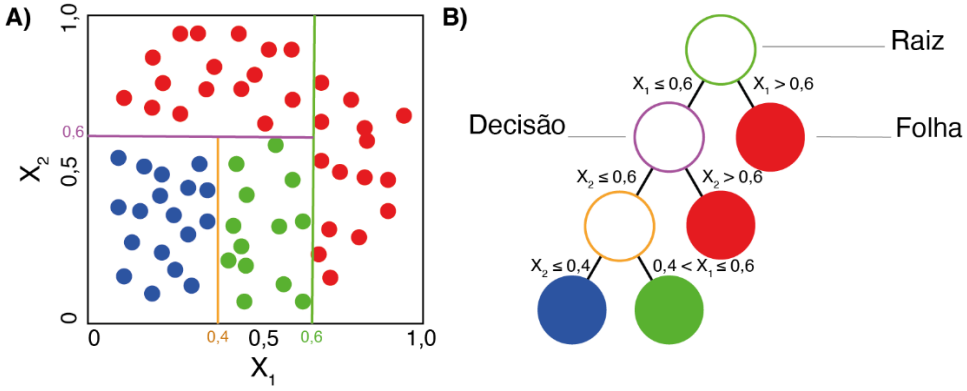
Fonte: (Wilimitis, 2018).

Para mais informações sobre o truque kernel, o autor recomenda a leitura de (Lou, Atoui e Li, 2021; Raschka e Mirjalili, 2019; Yan e Zhu, 2022). A partir de agora, será utilizada a sigla SVM para o modelo linear e kSVM para o algoritmo SVM com o truque do *kernel*.

3.3.4.4 Árvores de Decisão e Métodos de Combinação de Árvores de Decisão

As **árvores de decisão** são uma classe de algoritmos de aprendizagem de máquina não-paramétricos, bastante difundido no meio da análise de dados (Duda, Hart e Stork, 2000). Elas utilizam diversas questões, envolvendo estruturas lógicas, dividindo o conjunto de dados inicial em subconjuntos cada vez menores e, preferencialmente, puros, ou seja, que divida os exemplos em suas respectivas classes, como no exemplo da Figura 20 (Krzywinski e Altman, 2017; Raschka e Mirjalili, 2019).

Figura 20. Funcionamento da árvore de decisão.



Fonte: Autoria própria.

A árvore de decisão cria diversos hiperplanos no espaço amostral, como mostrado na Figura 20A, se tornando uma interessante alternativa para dados em que as classes não sejam linearmente separadas (Krzywinski e Altman, 2017). Uma árvore se inicia com o nó (todas as estruturas da árvore são chamadas de nós) raiz (Figura 20B), que possui todo o conjunto de dados de treinamento. Esse conjunto de dados é separado a partir de nós de decisão, que são questões a partir do qual o conjunto de dados pode ser dividido, que no caso da Figura 19, são limites no espaço das duas variáves X_1 e X_2. Porém, essa divisão pode ocorrer em diversas formas diferentes e a escolha de como serão gerados esses nós de decisão é de forma iterativa, geralmente utilizando algum tipo de otimizador, e é baseada em uma função objetivo (Raschka e Mirjalili, 2019). Uma dessas funções objetivo é o **ganho de informação** (IG, do inglês: *Information Gain*) (Equação 39).

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{39}$$

O ganho de informação é uma função da variável, f, que promoveu a separação e do subconjunto pai, D_p. N_j e N_p são o número de exemplos no j-ésimo nó filho e no nó pai, respectivamente. O I é uma medida de **impureza** dos nós. A partir da equação 39, pode se

observar que a IG é a subtração entre a impureza no nó pai e a soma das impurezas dos nós filhos. Porém, é importante agora definir o que é essa impureza (Duda, Hart e Stork, 2000).

Existem diversas formas de se calcular a impureza dos nós (Raschka e Mirjalili, 2019). Para o escopo desse trabalho, iremos focar apenas **impureza de Gini** (Equação 40), *G*.

$$G(t) = 1 - \sum_{k=1}^{K} p(k|t)^2 \tag{40}$$

Para essa equação tem-se: *K* é o número total de classes, *p(k|t)* é a proporção de elementos no nó *t* que pertencem à *k*-ésima classe. Nesse caso, temos a maior impureza em um nó, quando a proporção dos elementos de cada classe é igual, por exemplo, no caso de um problema binário (K = 2), tem-se: $\max G(t) = 1 - (0{,}5^2 + 0{,}5^2) = 0.5$. Já o mínimo é obtido quando se tem um nó puro, ou seja, há apenas elementos de uma classe no nó, nesse caso: $\min G(t) = 0$.

O objetivo final das árvores de decisão, a partir da estrutura até agora mostrada, é que todas os nós terminais, ou folhas, estejam puras, como apresentado na Figura 20. Porém, num caso real, ao chegar até o final de todas as divisões há uma grande possibilidade de a árvore possuir forte influência do ruído das amostras, podendo levar a um aumento da variância do modelo e, por fim, um sobreajuste do modelo (Raschka e Mirjalili, 2019).

Para evitar esse tipo de problema, na maioria dos casos, e nesse trabalho só será discutido esse tipo de processo, o momento de parada do algoritmo é configurada a partir de alguns hiperparâmetros como, o número máximo de divisões realizadas (a profundidade das árvores), valor mínimo do incremento da função objetivo entre um nó pai e um filho, ou o máximo de folhas geradas (Loh, 2011). Nesse caso, a escolha dos hiperparâmetros é uma parte essencial no treinamento desse tipo de algoritmo. Por conta disso, é recomendada uma otimização dos hiperparâmetros do modelo.

As árvores de decisão possuem sua principal vantagem atrelada ao fato da sua simplicidade e de poder ser uma alternativa para os modelos lineares, porém, sofrem com problemas de generalização, devido a heurística da busca pelas melhores divisões de nós, o que pode causar decisões baseadas em ótimos locais e serem enviesados a problemas com classes desbalanceadas (Raschka e Mirjalili, 2019).

Por conta desses contras, uma alternativa é combinar diversas árvores de decisão em um novo modelo mais robusto. Combinações entre estimadores mais fracos podem ser realizadas a partir de três processos conhecidos em inglês como ***voting***, ***bagging*** e ***boosting***. O

primeiro não fará parte do escopo do trabalho, para mais detalhes recomenda-se a leitura de (Raschka e Mirjalili, 2019). Inicialmente será abordado o método *bagging* (Figura 21).

Figura 21. Combinação de estimadores pelo método *bagging*.

Fonte: (Hachcham, 2023).

O método *bagging* realiza um sorteio (esse processo é conhecido em inglês como *bootstrap*), com reposição, entre os exemplos do conjunto de dados inicial, formando diversos novos subconjuntos de dados. Esses subconjuntos são treinados pelos classificadores e cada um deles realiza suas predições. Cada um dos classificadores vota e, para cada exemplo, são aplicadas as classes mais votadas por cada estimador (Raschka e Mirjalili, 2019; Zhao *et al.*, 2019).

Esse processo é o realizado para a construção das **florestas aleatórias** (RF, do inglês: *Random Forest*). A ideia por trás do algoritmo RF é a de melhorar a generalização das árvores de decisão a partir da criação de um modelo mais robusto, com a combinação de um número T de classificadores (Toth *et al.*, 2019; Zhao *et al.*, 2019). O modelo RF possui as etapas a seguir:

1. Construir o *bootstrap* de amostras, com cada subconjunto de dados possuindo n amostras.
2. Crescer cada árvore de decisão sendo que, para cada nó:
   a. Aleatoriamente selecionar f variáveis, sem reposição.
   b. Dividir o nó a partir da variável que proporciona os melhores valores da função objetivo.
3. Repetir os pontos 1 e 2 T vezes.
4. Juntar todas as árvores e a partir de maioria dos votos rotular cada uma das amostras.

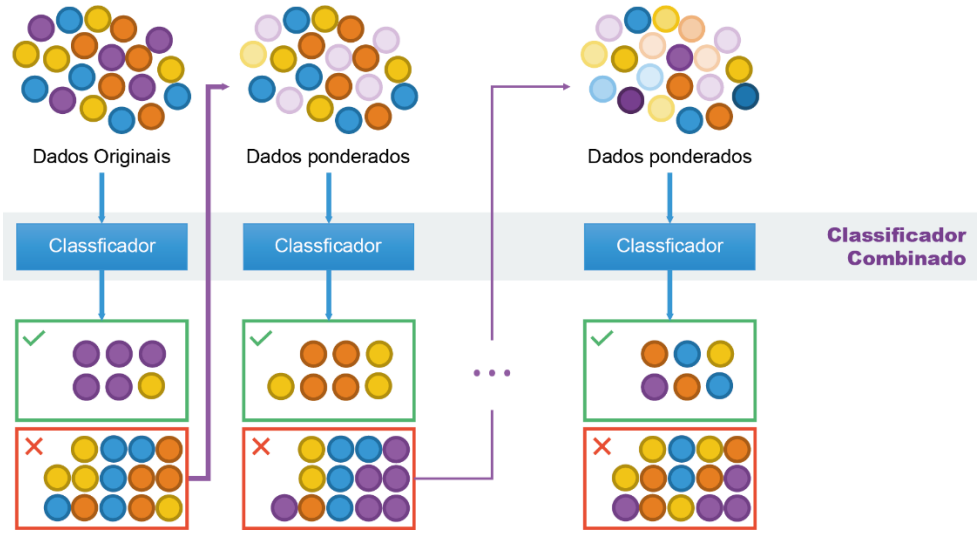
O algoritmo RF possui diversos hiperparâmetros a serem definidos previamente, por conta disso, recomenda-se a utilização de um método de otimização de hiperparâmetros

(Pedregosa, F. *et al.*, 2011). Um método derivado do RF é o algoritmo conhecido pelo nome inglês *Extremely Randomized Trees*, também conhecido como ***Extra-Tree*** (ETC). Esse algoritmo adiciona uma nova camada de aleatorieadade, pois, além de selecionar um subconjunto com *f* variáveis aleatórias, ainda randomiza também o limite de corte, que delimita a decisão a cada divisão nos nós (Geurts, Ernst e Wehenkel, 2006).

Já o método de combinação *boosting* (Figura 22) consiste nos seguintes passos chave:

1. Construir um subconjunto de treinamento com amostras aleatórias, d_1, sem reposição e treinar um classificador fraco, C_1.
2. Construir um segundo subconjunto de treinamento aleatório, d_2, sem reposição, adicionar uma porcentagem de amostras que foram classificadas erroneamente, e treinar um segundo classificador, C_2.
3. Treinar um terceiro classificador, C_3, a partir de um conjunto de dados, d_3, formado pelos exemplos que os classificadores C_1 e C_2 discordaram.
4. Combinar os três classificadores a partir de votação.

Figura 22. Combinação de estimadores pelo método *boosting*.



Fonte: (Hachcham, 2023).

Métodos combinados por *boosting* reduzem o viés, porém, podem criar a tendência de superajuste dos modelos (Raschka e Mirjalili, 2019). Como consequência da busca por melhoras desse tipo de combinação, foi criado o método de ***gradient boosting*** (Chen e Guestrin, 2017).

O *gradient boosting* possui os mesmos processos do *boosting* comum, porém a decisão de adicionar um novo classificador é realizado a partir de uma descida do gradiente para minimizar uma função de custo. Para mais detalhes sobre a descida de gradiente ver (Kwiatkowski, 2021).

Entre os membros da família de algoritmos que utilizam o gradiente boosting como combinação de classificadores débeis destacamos o ***Extreme Gradient Boosting*** (XGB) (Chen e Guestrin, 2016a). Como observado anteriormente, os métodos de boosting são métodos sequenciais, ou seja, cada classificador é adicionado após o anterior ser treinado. Esse fato reduz a velocidade de treinamento do modelo (Chen e Guestrin, 2016a). O algoritmo XGB, como uma solução para a redução desse tempo de execução, é implementado com uma paralelização na construção das árvores. Além disso, a arquitetura do algoritmo é escalável, ou seja, pode ser utilizado com uma grande variedade de conjunto de dados e implementado em diversas linguagens. Além disso, possui dois métodos de penalização que visa melhorar a generalização do modelo treinado (Chen e Guestrin, 2016a; Maheswara Rao *et al.*, 2023; Zhao *et al.*, 2020). Para mais detalhes sobre o algoritmo XGB ler (Chen e Guestrin, 2017).

3.3.4.5 K-Vizinhos Mais Próximos – KNN

O método KNN é um método computacionalmente simples que não cria uma função de decisão, mas, ao invés disso, computa todos os valores do conjunto de dados e faz sua classificação a partir da distância de cada exemplo às amostras vizinhas (Raschka e Mirjalili, 2019).

O algoritmo funciona de forma semelhante a uma CV, e segue os seguintes passos: (1) inicialmente uma amostra é retirada do conjunto de treinamento e (2) são computadas todas as distâncias (a distância Euclidiana, por exemplo) dessa amostra para as demais amostras do conjunto de treinamento. Uma vez que todas as distâncias foram computadas, (3) essas distâncias são dispostas em ordem decrescente. (4) Os *k*-exemplos com a menor distância, chamados de *k*-vizinhos mais próximos, são escolhidos para a votação e (5) a classe com mais votos, ou seja, a classe majoritária entre os vizinhos mais próximos é atribuída a essa amostra. No caso de empate, é somada distância entre todos os *k*-vizinhos de cada classe e a classe com menores distâncias é atribuída a amostra. Os cinco passos anteriores são repetidos para cada amostra até que, todas tenham sua classe atribuída pelo algoritmo (Ferreira, 2015; Raschka e Mirjalili, 2019).

Como pode ser visto, um atributo de grande importância na otimização do KNN é o valor do *k*. Esse é um hiperparâmetro que deve ser otimizado ao treinar esse tipo de modelo (Pedregosa, F. *et al.*, 2011).

Uma vez que, foram abordados os métodos de pré-tratamento, pré-processamento, e os algoritmos de classificação utilizados no presente trabalho, é importante citar quais as métricas utilizadas na avaliação da validação dos modelos de classificação.

3.3.4.6 Figuras de Mérito

A matriz de confusão ou tabela de contingência de um classificador (Quadro 3), torna possível calcular importantes figuras de mérito para a classificação. Na matriz de confusão, a classe verdadeira ou padrão ouro é comparada às classes estimadas pelo classificador, ou são comparados dois classificadores distintos. Por exemplo, tem-se duas classes: positivo, P, e negativo, N, nas colunas do Quadro 3 são apresentados os valores verdadeiros da classe e nas linhas os valores estimados pelo classificador.

Quadro 3. Exemplo de matriz de confusão ou tabela de contigência.

		Classe verdadeira		
		P	**N**	
Classe prevista	**P**	VP	FP	VP + FP
	N	FN	VN	FN + VN
		VP + FN	FP + VN	

Legenda: Classe P é a classe positiva, classe N é a classe negativa, VP são os verdadeiros postivos, FP são os falsos positivos, FN são os falsos negativos e VN são os verdadeiros negativos.
Fonte: Autoria Própria.

Analisando o Quadro 3, há quatro importantes valores, que são os números de amostras classificadas em cada um desses quadrantes. O VP, verdadeiro positivo, é o número de amostras que foram corretamente classificadas como positivas; o FP, falso positivo, é o número de amostras classificadas erradamente como positivas; o FN, falso negativo, é o número de amostras classificadas equivocadamente como negativas; e o VN, verdadeiro negativo, é o número de amostras classificadas acertadamente como negativas.

Em um exemplo de diagnóstico, pode se determinar H_0, a hipótese nula, como a ausência da doença e H_1, a hipótese alternativa, como a presença da doença, resultam disso a análise de dois tipos de erro: o erro tipo I, que seria tomar a decisão de classificar uma amostra como positiva quando ela não é, ou seja, um FP; e o erro tipo II, que seria tomar a decisão de classificar uma amostra como negativa, quando ela não é, ou seja, um FN.

Uma vez de posse da matriz de confusão, pode se calcular as figuras de mérito (Equações 41 – 43), Exatidão, Sensibilidade, Especificidade.

$$\text{Exatidão} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FN} + \text{FP} + \text{VN}} \tag{41}$$

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{42}$$

$$\text{Especificidade} = \frac{\text{VN}}{\text{FP} + \text{VN}} \tag{43}$$

O fator Kappa de Cohen (1960) foi criado como um coeficiente de concordância entre dois diferentes métodos classificadores. No seu trabalho, Cohen destaca duas proporções relevantes: p_o, a proporção de concordância entre os métodos (Equação 44), e p_e, a proporção de concordância ao acaso entre os métodos (Equação 45).

$$p_o = \frac{VP + VN}{N} \tag{44}$$

$$p_e = \frac{(VP + FP)}{N} \cdot \frac{(VP + FN)}{N} + \frac{(VN + FN)}{N} \cdot \frac{(VN + FP)}{N} \tag{45}$$

Nessas equações, o N representa o total de observações. Enfim, o Kappa de Cohen, κ, é calculado a partir da Equação 46 (Vach e Gerke, 2023).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{46}$$

Além dessas figuras de mérito citadas, uma importante medida de desempenho de um classificador binário é a curva Característica de Operação do Operador (ROC, do inglês: *Receiver Operator Characteristic*) e a área abaixo dessa curva (AUROC, do inglês: *Area Under the ROC curve*). A curva ROC é uma representação gráfica do deslocamento do limite entre as classes pelo classificador. Nesse gráfico, na abcissa são encontrados os valores da taxa de falso negativo, que equivale a "1 - Especificidade", já na ordenada, são encontrados os valores de Sensibilidade ou taxa de verdadeiro positivo. O limite superior da AUROC é 1, porém, um classificador que tenha essa área igual ou menor que 0,5 é considerado um classificador aleatório.

3.4 CÂNCER DE PRÓSTATA

O câncer de próstata (CaP) é a segunda neoplasia maligna mais frequente, depois do câncer de pulmão, em homens em todo o mundo, contando mais de um milhão de casos novos casos e causando mais de 300 mil mortes (6,8% de todas as mortes causadas por câncer em homens) em 2020 (Sung *et al.*, 2021).

A incidência e mortalidade do câncer de próstata aumentam com a idade. Homens negros possuem maior incidência que homens brancos. Outros fatores importantes na incidência e mortalidade de câncer de próstata são a dieta, prática de atividades físicas, tabagismo e uso de drogas (Rawla, 2019).

No Brasil, segundo o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) (2020), o câncer de próstata é o segundo tipo de câncer mais comum entre os homens, e é o segundo tipo de câncer com maior taxa de mortalidade em homens. Em 2018, o câncer de próstata foi responsável por mais de 15 mil óbitos, ficando atrás apenas do câncer de pulmão, traqueia e brônquios (Tabela 1).

Tabela 1. Mortalidade por câncer em 2018. Em destaque o câncer de próstata.

Localização	Óbitos	%
Traquéia, Brônquios e Pulmões	16.371	13,9
Próstata	**15.576**	**13,3**
Cólon e Reto	9.608	8,2
Estômago	9.387	8,0
Esôfago	6.756	5,8
Fígado e Vias Biliares Intrahepáticas	6.181	5,3
Pâncreas	5.497	4,7
Cavidade Oral	4.974	4,2
Sistema Nervoso Central	4.803	4,1
Laringe	3.859	3,3
Todas as Neoplasias	117.477	100,0

Fonte: (INCA, 2020).

O rastreamento do câncer de próstata, tema de muitas controvérsias, é a ferramenta que permite o diagnóstico nas fases mais incipientes da doença, onde a possibilidade de cura é maior que 90%. O método de rastreamento mais utilizado em todo mundo é o nível de antígeno prostático específico (PSA, do inglês *Prostatic Specific Antigen*) sérico, porém sofre de baixa especificidade.

Classicamente, a suspeita de câncer de próstata ocorre quando há uma alteração no exame físico (exame digital retal) e/ou nos valores séricos do PSA. No caso de ocorrência, deve-se proceder à biópsia do tecido prostático para confirmar o diagnóstico. A biópsia transretal é um exame desconfortável e doloroso. O paciente é posicionado desnudo, em decúbito lateral esquerdo, com flexão forçada dos membros inferiores. Após anestesia local, é então introduzido o probe ultrassonográfico, com preservativo não lubrificado e acoplado à

pistola de coleta de fragmentos. São coletados no mínimo 12 fragmentos de forma bilateral, mesmo quando há suspeita somente de um lado da próstata. A biópsia transretal possui complicações como sangramentos retais, hematúria e hematospermia[5], além de infecções locais (Araujo, 2016).

Uma vez que o câncer é diagnosticado, utiliza-se o escore da Sociedade Internacional de Patologia Urológica (ISUP, do inglês: *International Society of Urological Pathology*) e Gleason para estadiamento da doença, obtido a partir do estudo histopatológico do tecido prostático biopsiado. O escore ISUP-Gleason avalia o grau de diferenciação das células neoplásicas prostáticas, com escores variando entre 2 e 10. Quanto maior o valor, mais indiferenciado e agressivo é o tumor (ZEQUI, CAMPOS, 2010). Além disso, há o grupo de classificação patológica de Gleason, que leva em consideração o ISUP-Gleason e separa os pacientes entre os graus de 1 a 5, como referenciado na Tabela 2 (Kamitani *et al.*, 2021).

Tabela 2. Comparação entre os grupos de classificação de Gleason e o índice ISUP-Gleason.

Grupo de Classificação	ISUP-Gleason
1	≤ 6
2	7 (3+4)
3	7 (4+3)
4	8
5	9 e 10

Além da classificação ISUP-Gleason e do PSA, a categoria T patológica, pT, é um importante avaliador de risco do PCa. Ele é uma estimativa da extensão do tumor, a partir do tecido removido após a prostatectomia radical (Cho *et al.*, 2023). Além disso, a categoria pT, juntamente com a presença de margens cirúrgicas[6], é um importante critério na avaliação do risco de recidiva bioquímica do PCa (Diamand *et al.*, 2023; Remmers *et al.*, 2022).

O tratamento inicial pode ser na forma de prostatectomia radical, radioterapia e/ou terapia de privação de androgênio (ADT, do inglês *Androgen Deprivation Theraphy*), que se baseia no fato de que o crescimento do câncer de próstata depender do hormônio masculino predominante, a testosterona (Simon *et al.*, 2022). Após o tratamento, o teste de PSA pode ser empregado novamente para avaliar a recidiva da doença. A recidiva bioquímica é caracterizada pelo aumento dos níveis séricos de PSA pós prostatectomia radical e acomete até 40% dos pacientes pós prostatectomia radical (Cho *et al.*, 2023; Simon *et al.*, 2022).

[5] A hematúria e a hematospermia são o aparecimento de células sanguíneas, respectivamente, na urina e no esperma ejaculado.

[6] Margem cirúrgica positiva é caracterizado pela presença de células tumorais após a retirada da próstata, durante a prostatectomia. Para mais detalhes consultar (CHO et al., 2023).

No entanto, existem limitações bem documentadas em cada um desses estágios. Apenas cerca de 25% dos homens com nível elevado de PSA, definido como maior que 4,0 ng • mL^{-1}, são diagnosticados com câncer de próstata na biópsia e, inversamente, falsos negativos também são comuns (Kelly *et al.*, 2016). As biópsias frequentemente deixam de detectar o câncer devido à heterogeneidade do tumor, necessitando de múltiplas biópsias repetidas, que são potencialmente perigosas para os pacientes. As prostatectomias radicais estão associadas a comorbidades frequentes, incluindo disfunção erétil e incontinência urinária, mas estão associadas a melhores resultados de sobrevivência que a radioterapia. Da mesma forma, a ADT tem vários efeitos colaterais adversos e, em média, só é eficaz por dois a três anos antes do surgimento do câncer de próstata resistente à castração, uma doença incurável e muitas vezes fatal (Kelly *et al.*, 2016; Tourinho-Barbosa, Pompeo e Glina, 2016).

Diante de tudo isso, faz-se necessária a busca por novos métodos mais precisos com o objetivo de reduzir a procura por métodos de diagnósticos e estadiamento menos invasivos. Porém, a busca por novos métodos e novos biomarcadores está em fase de amadurecimento. Portanto, a metabonômica pode exibir vantagens tanto para a compreensão de doenças quanto para melhorar o diagnóstico e o monitoramento do tratamento, onde abordagens que se concentram em um único ou pequenos conjuntos de biomarcadores podem falhar em capturar a complexidade da patologia (Antcliffe e Gordon, 2016; Kelly *et al.*, 2016; Stabile *et al.*, 2020).

4 MATERIAIS E MÉTODOS

Nesse capítulo serão discutidas as etapas de obtenção das amostras, incluindo os critérios de inclusão e exclusão dos pacientes e voluntários envolvidos no estudo, as etapas de obtenção dos espectros, pré-processamento dos espectros para o conjunto de dados analisado, Conjunto de Dados de Diagnóstico do CaP, que teve como objetivo criar um modelo metabonômico baseado em RMN de ^{1}H para diferenciar pacientes saudáveis, daqueles com CaP. Por fim, num estudo preliminar (4.2), três conjuntos de dados foram separados a partir do Conjunto de Dados de Diagnóstico, a fim de avaliar o Estadiamento, o Risco de Recidiva Bioquímica e a ocorrência de Recidiva Bioquímica do CaP após a prostatectomia radical.

4.1 DIAGNÓSTICO DO CÂNCER DE PRÓSTATA

4.1.1 Amostragem

Foram estudados pacientes provenientes do Ambulatório de Urologia do Hospital das Clínicas da Universidade Federal de Pernambuco (HC-UFPE), com diagnóstico de câncer de próstata, bem como voluntários saudáveis, sem evidência de câncer de próstata.

Foram incluídos no estudo, pacientes do sexo masculino entre 40 e 75 anos de idade, com diagnóstico de câncer de próstata que foram submetidos à prostatectomia radical (Grupo PC) e pacientes sem evidências de câncer de próstata (Grupo NC).

Os critérios de inclusão do Grupo PC foram os seguintes:

- 40 a 75 anos de idade;
- Diagnóstico histológico de adenocarcinoma de próstata - USG sem evidência de outros tumores abdominais;
- Sumário de urina com menos de 5 hemácias por campo.

Os critérios de inclusão do grupo NC foram:

- 40 a 75 anos de idade;
- PSA < 1 ng/mL;
- Exame digital retal sem detecção de nódulos ou consistência alterada da próstata;
- USG sem evidência de outros tumores abdominais;
- Sumário de urina com menos de 5 hemácias por campo.

Os critérios de exclusão para ambos os grupos foram:

- Pacientes que fazem ou fizeram uso de inibidores da 5 alfa-redutase;
- Pacientes que fazem ou fizeram uso de bloqueadores androgênicos;

- Pacientes previamente submetidos à cirurgia prostática;
- Pacientes previamente submetidos à radioterapia pélvica;
- Pacientes com antecedente de qualquer neoplasia maligna (exceto próstata).

Os pacientes que fizeram ou estão fazendo uso de inibidores da 5 alfa-redutase, utilizados no tratamento da hiperplasia benigna, foram excluídos do teste pois, essas medicações provocam alterações no perfil metabólico, principalmente relacionados aos estrógenos, andrógenos e esteroides. Sobre mais informações sobre bloqueadores androgênicos o autor recomenda a leitura de (Lee *et al.*, 2020).

Os indivíduos que concordaram em participar do estudo, assinaram o Termo de Consentimento Livre Esclarecido (TCLE) e foram incluídos. A amostragem foi conduzida por conveniência, pois, uma vez respeitados os critérios de inclusão e exclusão, foram tomadas as amostras dos pacientes e voluntários que aceitaram os termos e assinaram o TCLE.

Foram incluídos 61 voluntários no estudo, sendo 38 pacientes diagnosticados com câncer de próstata. Todos os pacientes foram avaliados com exame digital retal, dosagem do PSA sérico, sumário de urina e ultrassonografia (USG) do abdome total e foram adicionados à classe PC. Os 23 pacientes que não acometidos por CaP foram adicionados ao Grupo NC, e tiveram amostras de sangue coletadas no ambulatório de urologia do HC - UFPE, após a realização da avaliação pré-determinada. Já os 39 pacientes do grupo PC, tiveram amostras coletadas na enfermaria de urologia do HC - UFPE, na véspera da realização da prostatectomia radical.

No consultório, foi realizado o processamento do sangue, através de centrifugação a 3000 rpm durante 5 minutos, sendo retirado o soro (sem os elementos figurados do sangue), que ficou armazenado a –20°C. As amostras congeladas foram encaminhadas ao Laboratório de Metabonômica e Quimiometria (LabMeQ) do Departamento de Química Fundamental (DQF) da UFPE, onde foram armazenadas em temperaturas inferiores a –40°C, para posterior análise por espectroscopia de RMN de ^{1}H na Central Analítica (CA) do DQF – UFPE.

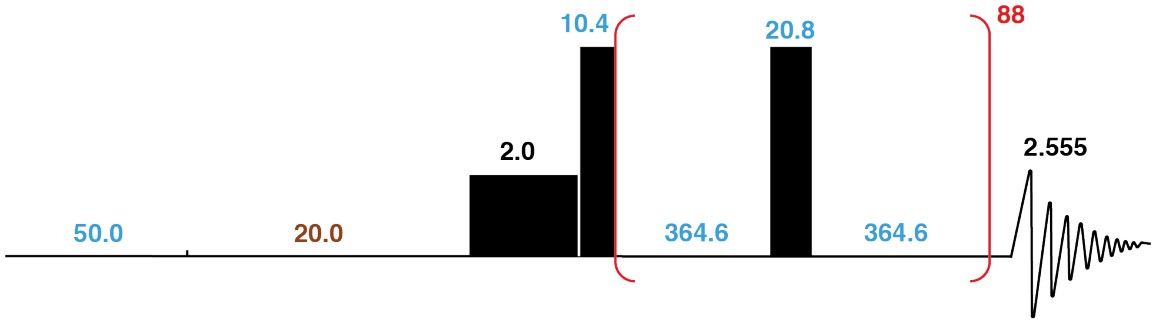
A comparação entre as médias dos dados demográficos foi realizada a partir do teste não paramétrico de Kruskal-Wallis (Lüpsen, 2023), para um nível de confiança de 95% realizada a partir do pacote *Scipy* (Virtanen *et al.*, 2020) em linguagem Python.

4.1.2 Espectrometria de RMN de ^1H

Os espectros de RMN de ^1H foram obtidos na Central Analítica do DQF – UFPE, em um equipamento da marca Agilent modelo VNMRS 400 de campo magnético de aproximadamente 9 T, a 400 MHz no canal de ^1H.

Uma vez descongeladas, 400 μL das amostras de soro foram adicionados a 200 μL de água deuterada (D_2O) e transferidos para tubos de RMN de 5 mm de diâmetro. Após a homogeneização da amostra diluída, o tubo foi inserido ao equipamento, então foi utilizada a sequência PRESAT/CPMG com a finalidade de suprimir o sinal da água e filtrar os sinais de substâncias de alta massa molar. Para os experimentos de RMN de ^1H do PRESAT foram utilizados os seguintes parâmetros: temperatura de 298 K, tempo de aquisição de 2,566 s, com quatro transientes. Para o experimento de PRESAT/CPMG tem-se: janela espectral de 4,8 kHz, temperatura de 298 K, tempo de aquisição de 2,566 s, 128 aquisições, tempo do pulso seletivo de PRESAT igual a 2,0 s, pulso de RF de 90°, τ = 0,4 ms e 88 ciclos entre os pulsos de RF de 180°, como descrito na sequência de pulsos na Figura 23.

Figura 23. Sequência de pulsos PRESAT/CPMG utilizada nesse trabalho.



Legenda: Preto, tempo em segundos; azul, tempo em microssegundos; marrom, tempo em milissegundos.

Foi realizado uma suavização do espectro utilizando a função *line broadening* de 0,3 Hz *software* OpenVNMRJ 3.1A integrado ao equipamento. O sinal da metila do lactato (δ 1,33 ppm) foi utilizado para referenciar o deslocamento químico do espectro obtido e a linha de base e distorções de fase foram corrigidos manualmente também utilizando o OpenVNMRJ 3.1A.

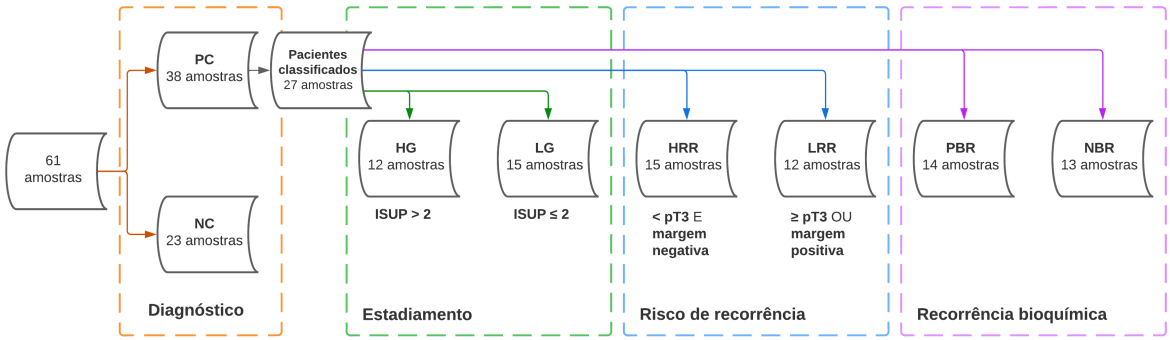
Após a aquisição dos espectros, foi utilizado o *software MestreNova 9.0*, da empresa *MestreLab Research,* onde foi utilizada a região entre δ 0,50 e 4,00 ppm, além disso, foram retiradas também as regiões entre δ 1,10 e 1,17 ppm; δ 1,90 e 1,93 ppm; δ 2,36 e 2,38 ppm, pois nesses intervalos foram encontrados picos anômalos que estavam presentes num número reduzido de amostras, o que poderia prejudicar a construção dos modelos estatísticos e suas posteriores interpretações. Além disso, com o mesmo aplicativo, foram realizadas a divisão dos

espectros em partes iguais, chamadas de *bins*, de 0,004 ppm de largura, resultando em um número total de 842 deslocamentos químicos – as variáveis da matriz de dados.

4.2 ESTUDO PRELIMINAR: ESTADIAMENTO, RISCO E OCORRÊNCIA DE RECIDIVA BIOQUÍMICA DO CÂNCER DE PRÓSTATA

O Conjunto de Dados de Estadiamento é um subconjunto do Conjunto de Dados Diagnóstico, apenas com os pacientes portadores de CaP. Dentre os 38 pacientes do grupo PC no Conjunto de diagnóstico, apenas 27 seguiram ao acompanhamento médico no ambulatório de urologia do HC-UFPE por, pelo menos, vinte e quatro meses após a prostatectomia radical (Figura 24).

Figura 24. Grupos do Conjunto de Dados de Diagnóstico, Risco de Recidiva, e de Recidiva Bioquímica.



Legenda: PC – com câncer de próstata; NC – sem câncer; HG – grau mais elevado de câncer de próstata (Gleason > 2); LG – grau menos elevado de câncer de próstata (Gleason ≤ 2); HRR – risco mais elevado de recidiva bioquímica (pT ≥ pT3 ou margem cirúrgica positiva); LRR – menor risco de recidiva bioquímica (pT ≤ pT2 e margem cirúrgica negativa); PBR – recidiva bioquímica positiva; NBR – recidiva bioquímica negativa.

Na Figura 24, são observadas as três etapas de classificação realizadas com os pacientes que foram acompanhados. A partir do exame da peça histopatológica, após a prostatectomia radical, foram realizadas as classificações do índice de ISUP-Gleason e a classificação histopatológica pT. No estadiamento, Conjunto de Dados 4, os quinze pacientes que possuíam o grupo de classificação de Gleason ≤ 2 foram dispostos no grupo de menor grau de CaP (LG) e os demais pacientes, grupo de classificação de Gleason > 2, foram dispostos no grupo de maior grau de CaP (HG). Num segundo momento, os pacientes foram classificados a cerca do risco de recidiva bioquímica a partir da classificação pT. Para o Conjunto de dados de risco de recidiva, doze pacientes estadiados como ≤ pT2 e margem cirúrgica negativa foram incluídos na classe de baixo risco de recidiva bioquímica (LRR), enquanto os quinze pacientes com a classificação histopatológica ≥ pT3 ou margem cirúrgica positiva foram incluídos como o grupo de alto risco de recidiva bioquímica (HRR). Por fim, com o objetivo de criar um modelo de prognóstico da recidiva bioquímica, o Conjunto de dados de recidiva bioquímica era

composto de treze pacientes, que durante os vinte e quatro meses de acompanhamento médico não fora observado a presença de recidiva bioquímica, foram classificados como grupo de recidiva bioquímica negativa (NBR), enquanto os catorze pacientes que apresentaram a recidiva bioquímica foram incluídos na classe positiva de recidiva bioquímica (PBR).

A seguir, será mostrado como foram realizadas as etapas quimiométricas com os seis conjuntos de dados.

4.3 ANÁLISE QUIMIOMÉTRICA

A etapa mediadora de qualquer modelo metabonômico é a quimiometria, uma vez que há informações químicas sendo desvendadas a partir de análise multivariada. Para tal, todas as etapas subsequentes utilizadas no presente trabalho – EDA, pré-processamento dos dados, visualização, seleção de variáveis e modelos de classificação, foram escritas em um documento *Jupyter Notebook* em linguagem *Python*.

Na Tabela 3 são apresentados os pacotes em linguagem Python, suas referências e suas propriedades na construção desse trabalho.

Tabela 3. Pacotes e bibliotecas Python utilizados no presente trabalho e suas principais funcionalidades.

Pacote	Propriedades	Ref.
Jupyter-lab	Ambiente de programação em que o desenvolvedor é capaz de testar trechos de código e ajustar como um relatório a partir dos *Jupyter Notebooks*.	(KLUYVER *et al.*, 2016)
Numpy	Permite a manipulação de vetores, matrizes e tensores. Além de incluir funções de álgebra linear, estatística básica etc.	(Harris *et al.*, 2020)
Pandas	Útil na manipulação de dados estruturados, importação e exportação de dados, entre outras funcionalidades.	(McKinney, 2010)
Scipy	Coleção de funções que permite a solução de problemas em álgebra linear, otimização, processamento de sinais, estatística, entre outros.	(Virtanen *et al.*, 2020)
Sklearn-genetic	GA para seleção de variáveis.	(Calzolari, 2022)
Scikit-learn	Possui diversas funcionalidades em ML. É dividido entre os módulos de métodos de classificação, regressão e agrupamento, redução de dimensionalidade, pré-processamento e avaliação e seleção de modelos (Validação cruzada, ajuste de hiperparâmetros, figuras de mérito etc.).	(Pedregosa, Fabian *et al.*, 2011)

Tabela 3. (Continuação)

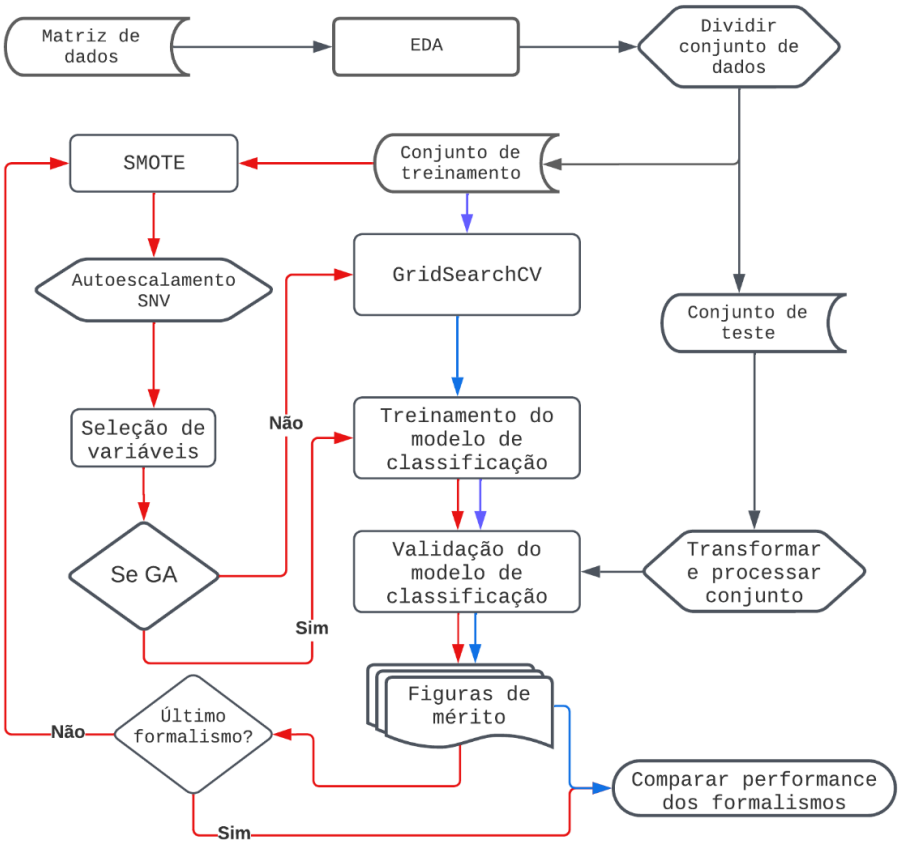
XGBoost	Implementa o método XGBoost para classificação e regressão.	(Chen e Guestrin, 2016b)
Imbalanced-learn	Possui diversas funções de reamostragem (sobreamostragem, subamostragem e métodos híbridos).	(Lemaitre, Nogueira e Aridas, 2016)
Matplotlib	Visualização de dados e imagens.	(Hunter, 2007)
Seaborn	Visualização de dados.	(Waskom, 2021)

Os conjuntos de dados, após serem normalizados utilizando o algoritmo SNV, foram autoescalados e então realizada uma EDA. Nessa EDA, se incluem a visualização dos espectros, visualização de dados após PCA e análise de amostras anômalas utilizando o T2 de Hotelling.

O **Conjunto de Dados de Diagnóstico** foi então separado em matrizes de treinamento e teste a partir do método de Kennard-Stone (KS) (Taylor, Kennard e Stone, 1969) com um tamanho do conjunto de teste de 30% do total de amostras da matriz de dados original para o grupo PC e de 50% para o grupo NC, uma vez que esse grupo tem um menor número de amostras.

Os conjuntos de treinamento, por sua vez, foram sobreamostrados na classe com menor número de amostras utilizando o algoritmo SMOTE. Em seguida, essas matrizes foram autoescaladas e suas médias e desvios padrão das variáveis utilizadas para autoescalar os conjuntos de teste. Após isso, finalmente ocorreu a estratégia de testar combinações entre métodos de seleção de variáveis e de classificação (Figura 25).

Figura 25. Fluxograma do processamento de dados realizado aos conjuntos de dados. As setas vermelhas indicam o laço realizado para a determinação do melhor formalismo a ser adotado no pré-processamento e classificação do conjunto de dados. Em azul, os processos realizados sem o SMOTE e a seleção de variáveis.



Sete diferentes tipos de seleção de variáveis mais uma etapa sem uma seleção de variáveis, foram colocadas num laço. Os métodos de seleção de variáveis são o método `SelectFromModel()`, do scikit-learn usando o classificador *ETC* (SFM-ETC) e o classificador LR (SFM-LR), um método de RFE com validação cruzada utilizando os classificadores *ETC* (RFECV-ETC) e LR (RFECV-LR), o método SFS, também com os classificadores *ETC* (SFS-ETC) e LR (SFS-LR) e o GA. Os métodos de seleção de variáveis foram confrontados com os algoritmos de classificação. Os estimadores testados foram o LDA, o SVM, o XGB, o KNN e o LR. Cada classificador, antes de ser treinado, com exceção do GA, passou por uma otimização dos hiperparâmetros utilizando o algoritmo `GridSearchCV()`. Os parâmetros buscados estão na Tabela 4. Os melhores estimadores foram então treinados. Além disso, para comparação, os classificadores também foram treinados utilizando o conjunto de dados sem o SMOTE e sem seleção de variáveis, linha em azul na Figura 25.

Tabela 4. Parâmetros utilizados na otimização dos hiperparâmetros de cada um dos algoritmos pelo método `GridSearchCV()`.

Parâmetros	**Valores para otimização**
SVM(1)	
regularização	[L1; L2]
C	[$1 \cdot 10^{-5}$ – 10]
LR(1)	
regularização	[L2; nenhuma]
C·	[$1 \cdot 10^{-5}$ – 10]
KNN(1)	
no. vizinhos	[2 – 7]
pesos	[uniforme; distância]
p	[1; 2; 3]
XGB(2)	
máx. profundidade	[3 – 18]
gama	[0 – 9]
regularização alfa	[0 – 80]
regularização lambda	[0 – 1]
taxa de aprendizagem	[0,01 – 0,3]
LDA(1)	
encolhimento	[sim; não]

Para mais detalhes sobre os parâmetros: (1) (Pedregosa, F. *et al.*, 2011) e (2) (CHEN e GUESTRIN, 2017).

Para avaliação do desempenho das combinações de seleção de variáveis e métodos de classificação, com e sem SMOTE, foram utilizadas cinco figuras de mérito, calculadas a partir da validação com o conjunto de dados de teste, quatro provenientes da matriz de confusão (Quadro 3), exatidão (Equação 36), sensibilidade (Equação 37), especificidade (Equação 38), o fator Kappa de Cohen (Equação 41), e a AUROC.

Os conjuntos: **Conjunto de Dados de Estadiamento**, **de Risco de Recidiva** e **de Recidiva** possuem número de amostras reduzidos, vinte e sete amostras, o que torna uma tarefa difícil realizar a modelagem. Para esses, foi realizada uma primeira etapa de separação dos grupos de treinamento e teste, utilizando o algoritmo KS, com o tamanho do conjunto de teste igual a 35% da matriz inicial. Após isso, o conjunto de treinamento foi sobreamostrado utilizando o algoritmo SMOTE, número de vizinhos igual a 4, que, por conta desse número reduzido de amostras, foi utilizado para reamostrar as duas classes. Para que, além de balancear o conjunto de dados, também fosse aumentado o conjunto final, as duas classes ficaram no final com o dobro do número de amostras da classe minoritária.

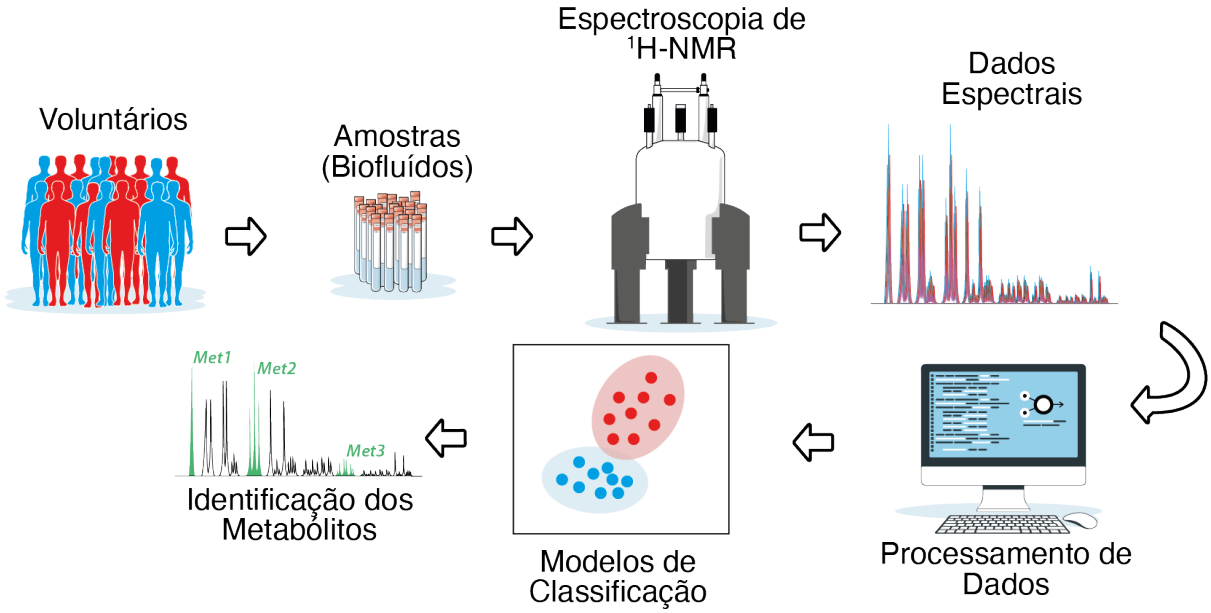
Após a etapa de reamostragem, utilizando a ferramenta de seleção de variáveis SFM-ETC, o conjunto de treinamento e teste tiveram sua dimensionalidade reduzida. Dentre os algoritmos árvore de decisão, SVM, LDA, KNN e LR, foi realizada uma busca manual pelo melhor modelo de classificação.

4.4 IDENTIFICAÇÃO E ANÁLISE DOS METABÓLITOS DE INTERESSE

Para o **Conjunto de Dados de Diagnóstico**, a partir das figuras de mérito, foi escolhido o melhor modelo para prosseguir com o fluxo de trabalho proposto. A partir do modelo treinado as variáveis mais importantes na classificação foram analisadas. A possível atribuição dessas variáveis foi realizada a partir da busca na literatura, a partir das variáveis de interesse dos modelos de classificação e utilizando o programa *NMR Suite 8.6*, da *Chenomx*, NW – Canadá.

As etapas do trabalho realizado para o **Conjunto de Dados de Diagnóstico** seguem o fluxo já consolidado no grupo de pesquisa (Figura 26), com a inserção de um laço para avaliar a melhor combinação entre seleção de variáveis e estimador, como já mostrado na Figura 25.

Figura 26. Fluxo comum de trabalho no grupo de metabonômica do LabMeQ/DQF/UFPE.



Fonte: Autoria própria.

5 RESULTADOS E DISCUSSÃO

Nesse capítulo, serão apresentados e discutidos os principais resultados obtidos, descrevendo, de forma sucinta, os caminhos percorridos durante a pesquisa científica. Esse trecho é organizado de acordo com o prosseguimento da ordem do método já estabelecido em nosso grupo de pesquisa, seguindo desde o perfil demográfico, passando pela análise exploratória, os modelos quimiométricos utilizados e a identificação dos metabólitos mais importantes na classificação para o **Conjunto de Dados de Diagnóstico**. Para os demais, serão mostrados os resultados e discussão acerca da análise exploratória e dos formalismos quimiométricos utilizados. Além disso, para demonstrar a capacidade de melhora da performance na generalização de modelos quimiométricos com a sobreamostragem e a seleção de variáveis, foram utilizados dois conjuntos de dados metabolômicos baseados em RMN de ^{1}H, já publicados, Conjunto de Dados de Diagnóstico de Asma em Gatos (5.3) e de Imunização de Varíola (5.4), e encontrados no *Metabolomics Workbench* (Sud *et al.*, 2016).

5.1 DIAGNÓSTICO DE CaP

As características demográficas e clínicas são mostradas abaixo na Tabela 5.

Tabela 5. Características demográficas dos participantes do Grupo 1 e do Grupo 2.

	Grupo 1 Mediana (Q1 – Q3)(1)	**Grupo 2** Mediana (Q1 – Q3)	**Significância estatística**(2)
Número de indivíduos	23	39	
Idade	49,5 (44,5 – 58.5)	70,0 (66,0 – 74,0)	< **0,01**
PSA (ng · L^{-1})	0,58 (0,39 – 0,69)	8,03 (5,76 – 14,4)	< **0,01**
ISUP escore Gleason		Número de indivíduos	
1	-	3	
2	-	20	
3	-	13	
4.....	-	0	
5	-	2	

(1) Q1 e Q3 são o 1º e 3º quartis, respectivamente.
(2) O p-valor foi medido a partir do teste de Kruskal-Wallis.

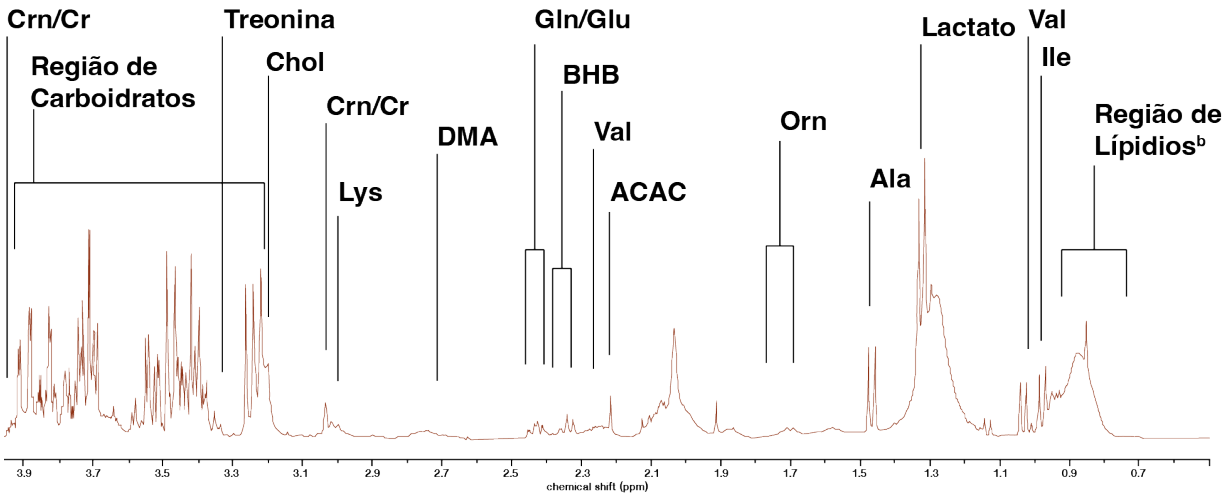
Na Tabela 5 são verificadas que há claramente diferença estatística entre as médias populacionais ($p < 0,05$) no que pesam as características dispostas – idade e PSA. Isso é esperado para o PSA, uma vez que é utilizada na detecção clínica do CaP. No caso das idades, a diferença entre os grupos foi atribuída à amostragem por conveniência, desde que a aquisição de amostras depende da demanda do ambulatório. Ademais, a prevalência do câncer de próstata

aumenta com a idade, sendo um dos motivos da não equidade das médias entre os grupos. Outro fator digno de nota é o valor do PSA < 1 ng · mL^{-1} para inclusão no Grupo 2 do estudo, sendo importante salientar que, os valores de PSA e do volume prostático também aumentam com a idade, explicando a faixa etária mais baixa no Grupo 2 (Gustafsson e Mansour, 1998).

Essa questão poderia ser um fator limitante na análise desses dados, uma vez que esta informação pode estar incorporada a variância dos dados. Porém, como os formalismos analisados foram supervisionados esses outros fatores não têm influência no modelo, uma vez que esse não possui correlação direta com o vetor resposta.

A seguir, será abordada a análise dos dados obtidos após a análise por RMN. Os espectros obtidos possuem um perfil como indicado na Figura 27, sendo interessante frisar que as amostras possuem o mínimo de manipulação, não sendo necessário um preparo de amostras mais rebuscado, apenas a diluição em água deuterada, o que torna uma análise mais simples e com menor propensão a erros sistemáticos, advindos dessa etapa.

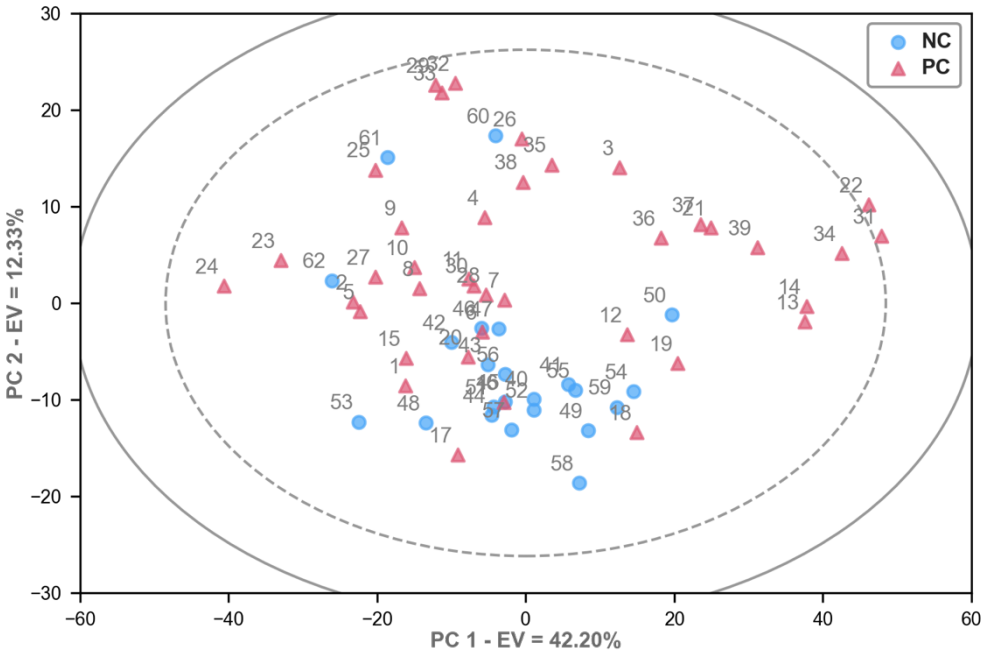
Figura 27. Espectro de RMN de ^{1}H (PRESAT/CPMG, 400 MHz) de uma das amostras de soro de sangue, após a retirada das regiões maiores que 4 ppm e menores que 0,5 ppm. Em destaque alguns metabólitos de interesse[a].



a. Os metabólitos indicados foram obtidos a partir do artigo de Gowda e colaboradores (2015) e do aplicativo *NMR Suite 8,6* (versão de avaliação), da *Chenomx* (NW, Canadá).
b. Região das metilas de lipídios, predominantemente LDL e VLDL.
Abreviações: ACAC Acetoacetato; Ala – Alanina; BHB b-hidroxibutirato; Chol Colina; Crn/Cr Creatinina/Creatina; DMA Dimetilamina; Gln/Glu Glutamina/Glutamato; Ile Isoleucina; Lys Lisina; Orn Ornitina; Val Valina.

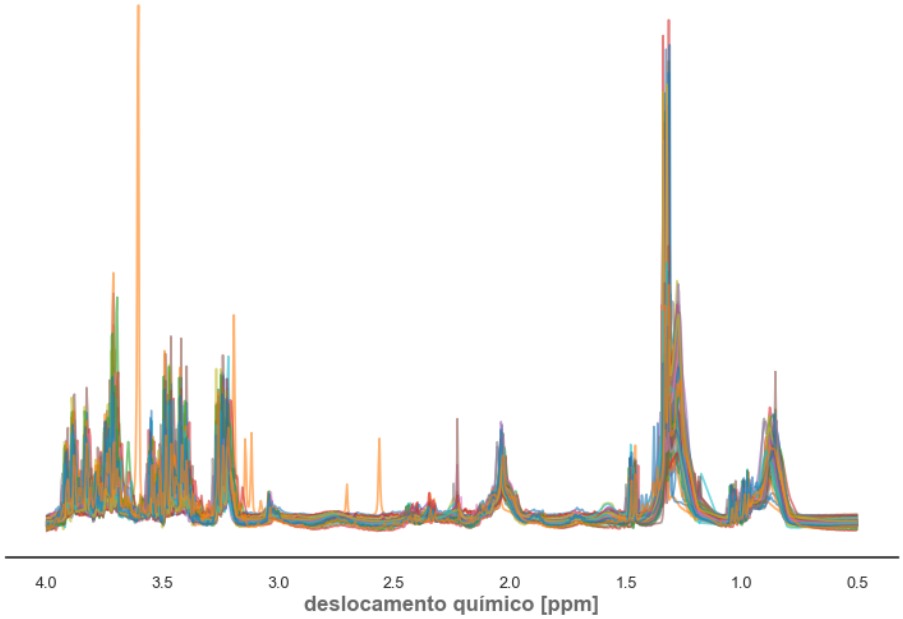
Uma vez obtidos e pré-processados os espectros, obtemos uma matriz de dados com 62 linhas (amostras) e 842 deslocamentos químicos (variáveis). Com a matriz de dados em mãos, a primeira etapa é a de EDA, utilizando a PCA (Figura 28).

Figura 28. Escores da PCA do Conjunto de dados de diagnóstico. Adicionalmente são mostradas as elipses de T^2 de Hotelling de 95% e 99% de confiança.



Como pode ser observado na Figura 27, há duas amostras ligeiramente fora da elipse de 95% de confiança de T2 de Hotelling, as amostras 22 e 31. Fazendo uma inspeção visual dessas amostras (Figura 29), nota-se que a amostra 22 se destaca em meio aos demais espectros. Então, foi concluída que ela estava fora das especificações das demais amostras e eliminada do estudo. A amostra 31 foi mantida, uma vez que a distância dela para o limite era pequena e o número de amostras já é bastante reduzido.

Figura 29. Sobreposição dos espectros de RMN de ^{1}H das amostras do Conjunto de dados de diagnóstico.

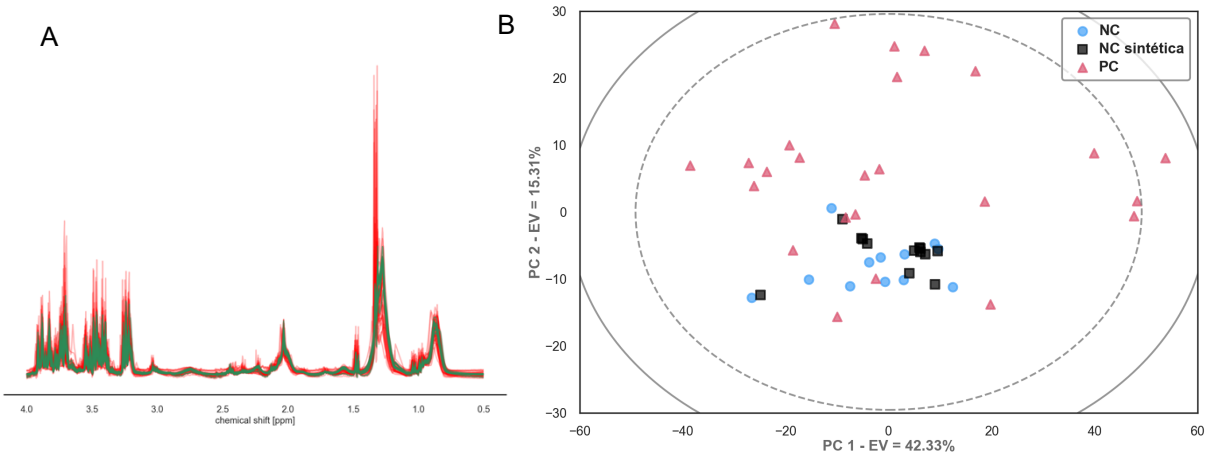


Ainda em relação à PCA (Figura 28), a partir desse método não-supervisionado não há nenhuma tendência de separação entre as classes de interesse.

Após eliminar a amostra 22, a matriz inicial foi dividida em dois subconjuntos, um para o treinamento dos modelos (PC = 25 amostras, NC = 11 amostras) e um para a validação externa (PC = 13 amostras, NC = 12 amostras). O conjunto de treinamento foi autoescalado e seus dados de média e desvio padrão utilizados para o autoescalamento do conjunto de teste. Uma vez que, passaram por esse pré-processamento, o conjunto de treinamento foi utilizado para construir cinco modelos de classificação.

Sucessivamente à construção dos modelos de classificação sem reamostragem e sem seleção de variáveis, seguiu-se à próxima etapa, que é o laço contendo as seleções de variáveis e estimadores. Antes disso, o conjunto de treinamento foi reamostrado utilizando o algoritmo SMOTE. Com esse algoritmo foi igualado o número de amostras dos dois grupos, ou seja, ambas as classes, PC e NC, possuíam agora 25 amostras no grupo de treinamento. Na Figura 30 são apresentados os escores da PCA e os espectros dos dados pós SMOTE, destacando as amostras que foram geradas**.**

Figura 30. EDA dos dados após SMOTE para o Conjunto de dados de diagnóstico. A) Os espectros em verde são das amostras geradas por SMOTE. B) Escores da PCA destacando as amostras sintéticas.

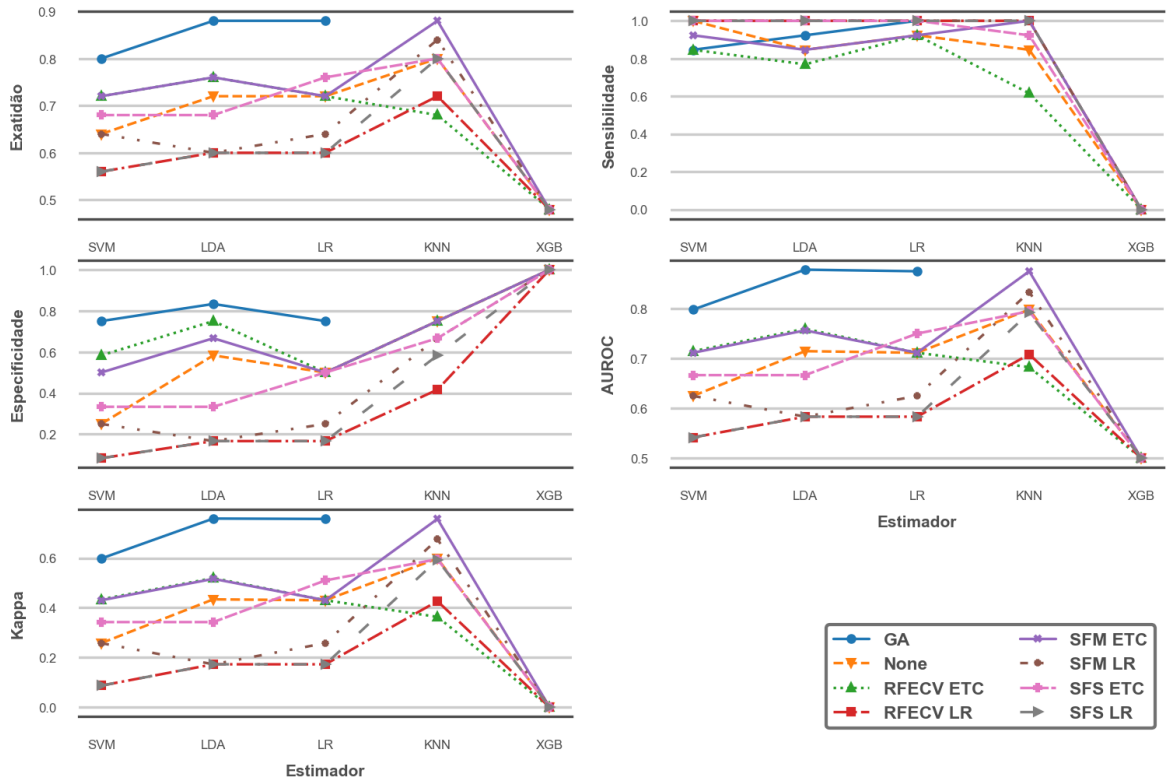


Na Figura 30A, nota-se que as amostras criadas a partir do SMOTE, em verde, seguem um perfil parecido com um espectro de RMN, assim como as amostras originais. Essa inspeção é importante para não adicionar informações irrelevantes para a fase de análise das variáveis de interesse. Na Figura 30B, pode ser observada a distribuição das amostras sintéticas mais próximas do centro da distribuição das amostras. Por ser um método que "povoa" o hiperespaço de amostras com novos exemplos em posições aleatórias, apenas relativas ao número de vizinhos próximos, o SMOTE pode gerar amostras de fronteira dificultando a criação de modelos de classificação ou pode gerar amostras próximas ao centro de distribuição dos dados,

não considerando o ruído e podendo causar superajuste do modelo, tornando a escolha do número de vizinhos levados em consideração pelo modelo uma parte fundamental na sobreamostragem. A escolha de um menor número de vizinhos pode reduzir essa marginalização, porém, há também o risco de sobreajuste, uma vez que o SMOTE também não adiciona ruído às amostras sintéticas, como já citado anteriormente (Beinecke e Heider, 2021; Shen *et al.*, 2021; Sreejith, Khanna Nehemiah e Kannan, 2020; Wang *et al.*, 2021).

O novo conjunto de treinamento de dimensão (50 × 842) foi autoescalado e, da mesma forma que anteriormente, o conjunto de teste. Após o pré-processamento, o conjunto de treinamento foi utilizado no laço para criar os modelos de classificação com cada um dos métodos de seleção de variáveis. Todos os modelos tiveram seus hiperparâmetros otimizados a partir do método `GridSearchCV()`, com exceção dos algoritmos de classificação após GA, já que não há como otimizar os hiperparâmetros e ao mesmo tempo realizar a seleção de variáveis. O melhor modelo foi então treinado e o conjunto de teste foi utilizado para obter as figuras de mérito mostradas no gráfico abaixo (Figura 31) e com mais detalhes no **Apêndice B**.

Figura 31. Resultado dos formalismos propostos para o Conjunto de dados de diagnóstico.



Na Figura 31, fica claro que os melhores resultados obtidos durante a validação são obtidos após a seleção de variáveis com GA e a combinação SFM-ETC-KNN. O GA é um algoritmo meta-heurístico com grande custo computacional, mesmo sendo um algoritmo considerado simples. Uma vez que, ele é extremamente dependente dos hiperparâmetros

selecionados *a priori*, a escolha desses é uma etapa crucial no desempenho do método de seleção de variáveis. Nesse caso, foi utilizado um valor de população de 700 e um valor de iterações ou gerações de 1000 para ser avaliadas. Um incremento nesses valores, por exemplo, pode garantir que o GA alcance melhores valores, já que ele vai ter um maior espaço amostral para analisar. Porém, o GA possui a desvantagem de no seu processo de otimização, poder alcançar apenas máximos locais (Guha *et al.*, 2021). O GA é um algoritmo meta-heurístico, ou seja, ele busca otimizar o melhor conjunto de dados, mas como não é um método de "força bruta", não é capaz de analisar todas as soluções possíveis, o que seria muito custoso computacionalmente (Galvão, Araújo, de e Soares, 2020). Porém, nesse caso, o GA encontrou os conjuntos de dados mais apropriados para o SVM, LDA e LR, nesse caso específico.

O método SFS, na busca por 24 variáveis, possui valores de AUROC e Kappa abaixo ou compatíveis com os resultados sem nenhuma seleção de variáveis. Porém, o SFS, assim como o GA são algoritmos com hiperparâmetros importantes que podem ser ajustados. No entanto, fazer uma otimização desses fatores teria um custo computacional grande, uma vez que envolve muitas iterações e um volume grande de dados na memória do computador. Para o SFS, um número de variáveis selecionadas arbitrário foi escolhido. A Tabela 6 apresenta os resultados da validação para os dados com e sem SMOTE e sem seleção de variáveis.

Tabela 6. Figuras de mérito dos modelos sem uso do SMOTE. Entre parênteses, estão os valores com emprego de SMOTE.

Estimador	Exatidão	Sensibilidade	Especificidade	AUROC	Kappa
SVM	0,80(0,64)	0,85(1,00)	0,75(0,25)	0,80(0,63)	0,60(0,26)
LDA	0,72(0,72)	0,85(0,85)	0,58(0,58)	0,72(0,72)	0,43(0,43)
LR	0,72(0,72)	0,92(0,92)	0,50(0,50)	0,71(0,71)	0,43(0,43)
XGB	0,60(0,52)	0,85(1,00)	0,33(0,00)	0,59(0,50)	0,18(0,00)
KNN	0,80(0,80)	0,85(0,85)	0,75(0,75)	0,80(0,80)	0,60(0,60)

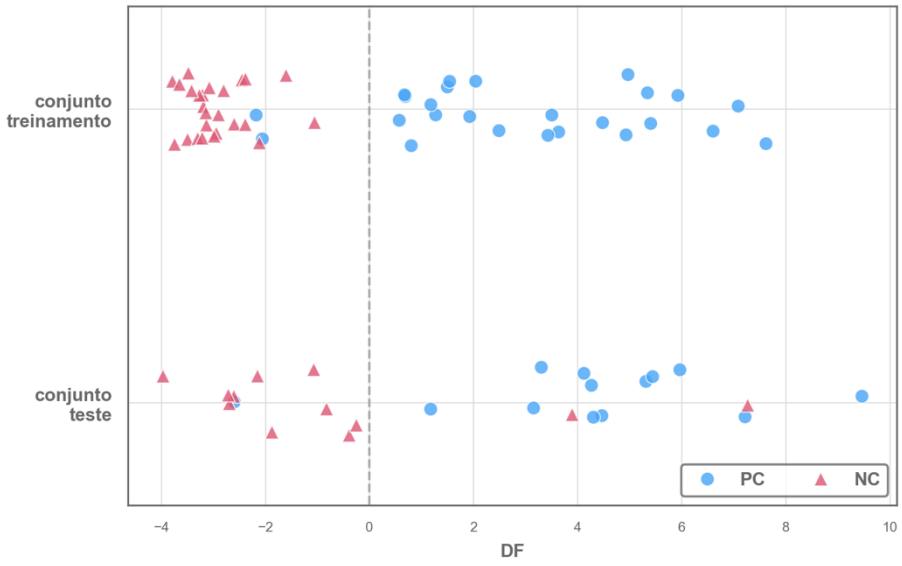
Abreviações: SVM – Máquina de Vetores de Suporte; LDA – Análise Discriminante Linear; LR – Regressão Logística; XGB – *Extreme Gradient Boost*; KNN – K-Ésimo Vizinho mais próximo, AUROC – Área Abaixo da Curva Característica do Operador.

Como exibido na Tabela 6, entre os modelos construídos sem emprego do SMOTE e da seleção de variáveis durante o pré-processamento, aqueles que apresentaram melhor desempenho são o SVM e o KNN. Adicionalmente, não há melhora nos valores apenas com o uso de SMOTE, valores entre parênteses. Mesmo com a geração de novas amostras do grupo NC, os modelos continuam enviesados para o grupo com maior número de amostras reais, pois os valores de sensibilidade são sempre maiores que os da especificidade, nos modelos com emprego de SMOTE.

Objetivando a escolha do melhor formalismo e pré-processamento para o problema proposto, utilizou-se a AUROC e o fator Kappa de Cohen como as figuras de mérito mais importantes na classificação. Para os formalismos GA-LDA, GA-LR e SFM-ETC-KNN, tivemos, respectivamente, os valores de AUROC de 0,878, 0,875 e 0,875 e para o Kappa de 0,759, 0,757 e 0,757. Nesse caso, o LDA é um algoritmo que gera uma visualização mais simples, pois também promove uma redução da dimensionalidade e ainda permite a possibilidade de análise direta da importância das variáveis (Neto *et al.*, 2020). Essa última informação também é válida para o modelo de LR, porém o KNN é um modelo "caixa preta", ou seja, não fornece quaisquer possibilidades de obtenção das variáveis mais importantes na classificação (Raschka e Mirjalili, 2019). Por conta disso, o modelo metabonômico utilizado nesse trabalho foi o GA-LDA.

Com as variáveis selecionadas pelo GA, o modelo LDA construído, utilizando um método de penalização não apresentou sobreajuste, pois os valores de exatidão do modelo no treinamento e no teste foram bem próximos, como pode ser observado na matriz de confusão (Quadro 4) e no gráfico de escores dos conjuntos de teste e treinamento (Figura 32).

Figura 32. Função discriminante da LDA para os conjuntos de teste e treinamento do Conjunto de dados de diagnóstico.



Quadro 4. Matriz de confusão do modelo GA-LDA. Em negrito, estão os valores de VP e VN.

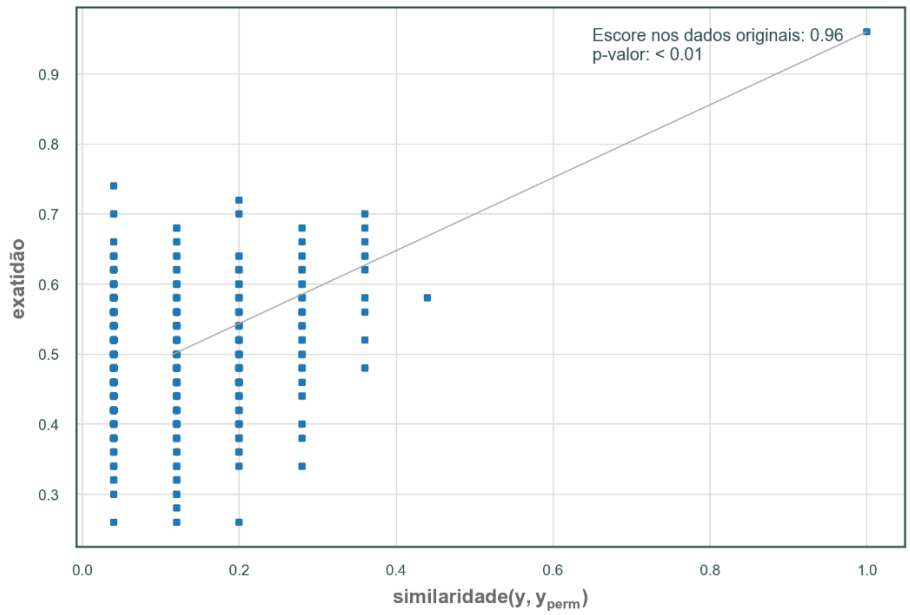
Modelo GA-LDA	**Diagnóstico Clínico**		
	PC	*NC*	*Total*
PC	**23(12)**	0(2)	23(14)
NC	2(1)	**25(10)**	27(11)
Total	25(13)	25(12)	

No conjunto de validação, o modelo possui um falso negativo e dois falsos positivos, apresentando valores de sensibilidade e especificidade de 92% e 83%, além de uma exatidão

de 88%. No conjunto de treinamento, as amostras sintéticas geradas reduziram a dispersão dos dados da classe NC, esse tipo de comportamento da reamostragem pode acarretar problemas de generalização no conjunto de validação, por isso, é essencial o cuidado em quantas amostras serão geradas e o método adequado de sobreamostragem (Beinecke e Heider, 2021; Rodrigues, Luna e Pinto, 2023).

Além disso, um teste de permutação foi realizado com a exatidão na validação cruzada 5-*fold* (Figura 33).

Figura 33. Teste de permutação do GA-LDA para o Conjunto de dados de diagnóstico. O p-valor é dado pelo número de valores de exatidão com os dados permutados maior que com os dados originais dividido pelo número total de permutações.



Considerando todas as figuras de mérito acima e a significância estatística do modelo (Figura 33), seguiu-se a análise das variáveis de importância na classificação. Essa etapa é de grande valia para a metabonômica, pois os metabólitos de interesse na classificação podem gerar importantes contribuições no entendimento da doença (Nicholson e Lindon, 2008).

As variáveis selecionadas pelo GA com o estimador LDA foram: δ 1,34 ppm, δ 1,70, δ 1,98 ppm, δ 2,26 ppm e δ 3,87 ppm. A partir do Chenomx e dois artigos, esses deslocamentos químicos foram atribuídos aos seguintes metabólitos, ou conjunto de metabólitos, respectivamente: lactato, ornitina, lipídios, valina e carboidratos (Nagana Gowda, Gowda e Raftery, 2015; Nicholson *et al.*, 1995). A distribuição dessas variáveis com a classe de interesse se encontra no gráfico de caixas na Figura 34, enquanto o gráfico dos coeficientes do GA-LDA se encontra na Figura 35.

Figura 34. Distribuição das variáveis selecionadas pelo GA para o classificador LDA.

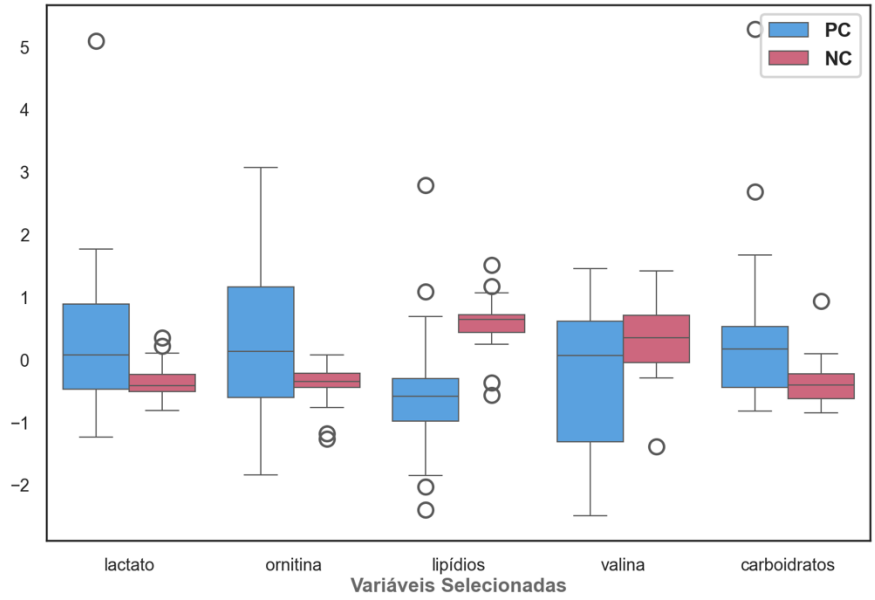
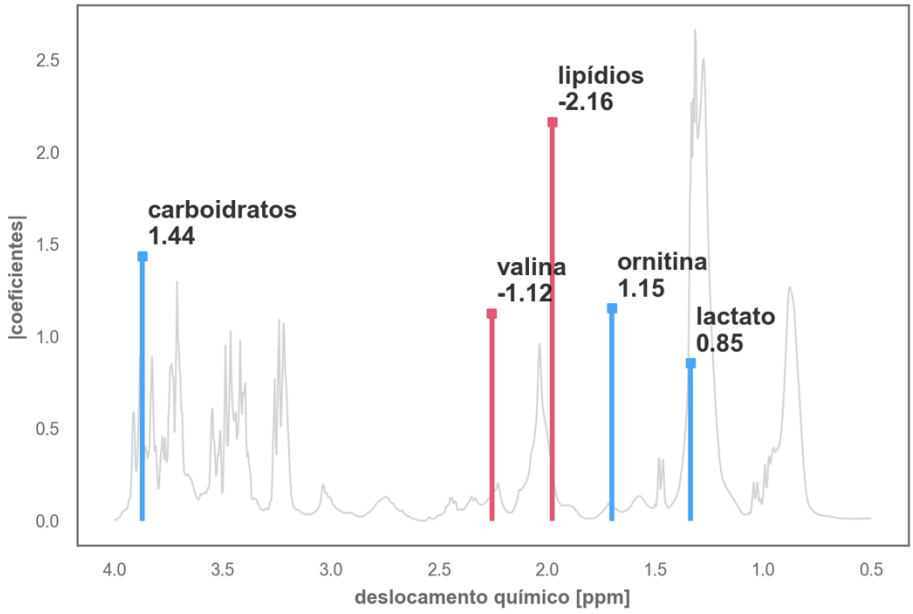


Figura 35. Coeficientes lineares do GA-LDA para o Conjunto de dados de diagnóstico.



Nesse caso, os valores positivos indicados no gráfico da Figura 35 são dos metabólitos que estão expressos no soro em maior quantidade em pacientes com CaP, que são os casos de ornitina, lactato e de carboidratos. Enquanto aqueles com valores negativos, valina e os lipídios, estão em maiores quantidades nas amostras de pacientes do grupo NC.

Segundo Zheng et al, (2020), em células normais, uma boa parte da demanda energética é suprida pelo ciclo do citrato (TCA). Porém, em células cancerígenas, devido à grande demanda de energia para o crescimento do tumor, essas células buscam outras fontes. Um exemplo é suprir essa demanda a partir da glicólise aeróbia, em que a glicose é convertida

a lactato, também conhecida como Efeito de Warburg (Bueno De Paiva *et al.*, 2021). Por conta da limitação dos níveis de glicose, gerado pelo consumo excessivo, a célula tumoral inicia o processo de produção de glicose, a partir de substratos que não são carboidratos. Esse processo chamado de gliconeogênese é responsável pela síntese de metabólitos importantes no crescimento do tumor (Grasmann *et al.*, 2019). O trabalho de Wang e Dong (2019) mostra que existe uma regulação positiva nos níveis da enzima responsável pelo controle do fluxo gliconeogênico durante a multiplicação de células neoplásticas de CaP. Essa expressão positiva de glicose é o que pode explicar os valores positivos dos coeficientes do GA-LDA.

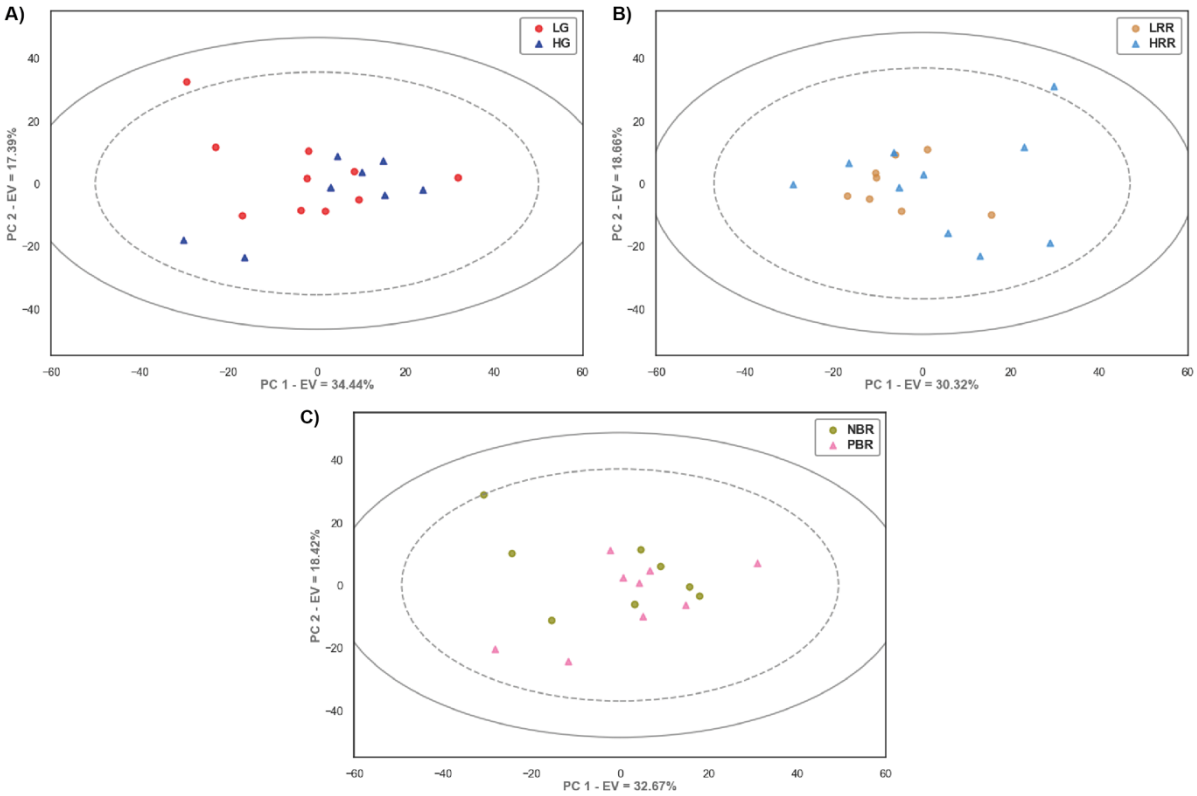
Quanto aos lipídios, que aparecem com coeficientes negativos, a sua perturbação pode ser explicada também pela alta demanda energética da proliferação desordenada das células neoplásticas. Xu et al (2021) indicam que o aumento do consumo de algumas espécies de lipídios pode ser causado pela gênese de membrana durante a carcinogênese, causando a redução dos níveis séricos dos lipídios.

A perturbação nas expressões dos aminoácidos pode ser explicada pela requisição destes como fonte de energia em caso de carência de carboidratos ou lipídios. Um outro ponto que pode afetar na desregulação da concentração de aminoácidos é quanto a necessidade de blocos construtores para o crescimento tumoral (Zheng *et al.*, 2020). Os achados desse estudo foram publicados em (Oliveira *et al.*, 2023).

5.2 CONJUNTOS DE DADOS DE ESTADIAMENTO, RISCO E OCORRÊNCIA DE RECIDIVA BIOQUÍMICA DO CÂNCER DE PRÓSTATA

Os Conjunto de Estadiamento, Risco de Recidiva e Recidiva Bioquímica foram primeiramente submetidos a uma PCA, para avaliar se havia algum agrupamento na técnica não-supervisionada ou a presença de amostras anômalas (Figura 36).

Figura 36. Escores das PCA realizadas para (A) Conjunto de dados de estadiamento, (B) Conjunto de dados de risco de recidiva e (C) Conjunto de dados de recidiva bioquímica.



A partir dos escores da PCA (Figura 36), não pôde ser verificada a presença de algum agrupamento e, foi verificada apenas a presença de amostras extremas, mas nenhuma amostra anômala.

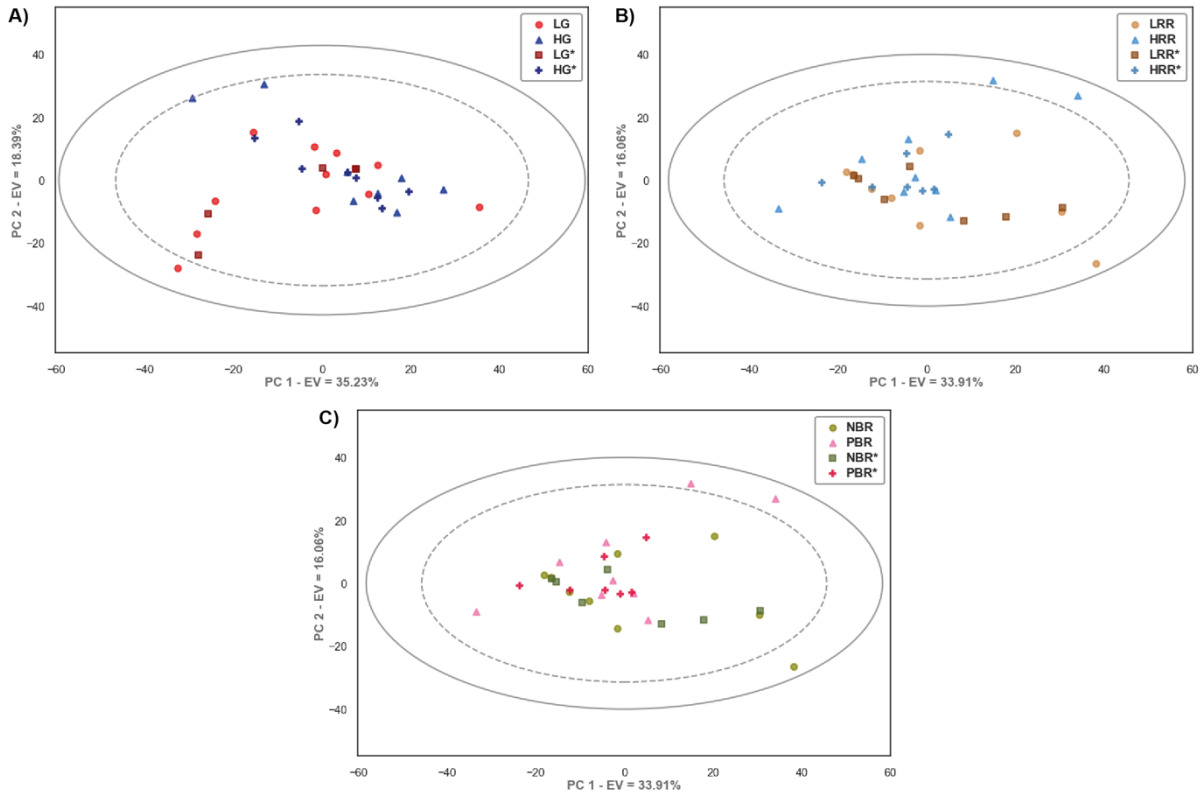
Após separar os conjuntos de dados em conjuntos de treinamento e teste, observou-se um número reduzido de amostras (Tabela 7). Por conta disso, os conjuntos de treinamento resultantes da divisão de Conjunto de Dados de Estadiamento, Conjunto de Dados de Risco de Recidiva e Conjunto de Dados de Recidiva Bioquímica foram reamostrados nas duas classes, como mostrado na Tabela 7.

Tabela 7. Distribuição do conjunto de teste e dos conjuntos de treinamento antes e depois do SMOTE. As distribuições para cada conjunto são: Conjunto de dados de estadiamento = [LG, HG]; Conjunto de dados de risco de recidiva = [LRR, HRR]; e Conjunto de dados de recidiva bioquímica = [NBR, PBR].

Conjunto de Dados	Conjunto teste	Conjunto trein. pré SMOTE	Conjunto trein. pós SMOTE
Estadiamento	[5, 4]	[10, 8]	[16, 16]
Risco de recidiva	[4, 5]	[8, 10]	[16, 16]
Recidiva bioquímica	[5, 5]	[8, 9]	[16, 16]

Os conjuntos de treinamentos resultantes foram avaliados a partir de uma PCA (Figura 37) e avaliados os seus perfis.

Figura 37. Escores das PCA realizadas para a (A) Conjunto de dados de estadiamento, (B) Conjunto de dados de risco de recidiva e (C) Conjunto de dados de recidiva bioquímica, após o SMOTE. Nas legendas, os valores com asterisco são das amostras sintéticas para cada grupo.



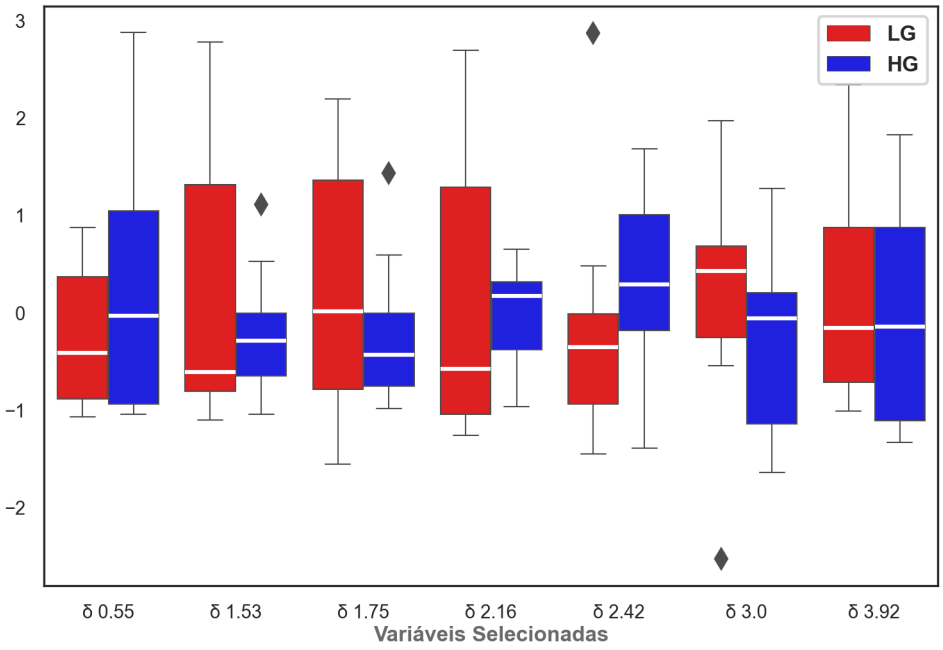
Mais uma vez, como mostrado na Figura 37, o SMOTE, por ser um método que usa os vizinhos mais próximos para criar amostras sintéticas, concentra sua geração mais próximo do centro da distribuição das amostras, reduzindo a dispersão do conjunto de dados (Rodrigues, Luna e Pinto, 2023).

A partir desse ponto, para melhorar o entendimento do texto nesse trecho do trabalho, esse item será separado a partir das três etapas: estadiamento, risco de recidiva bioquímica e recidiva bioquímica.

5.2.1 Estadiamento

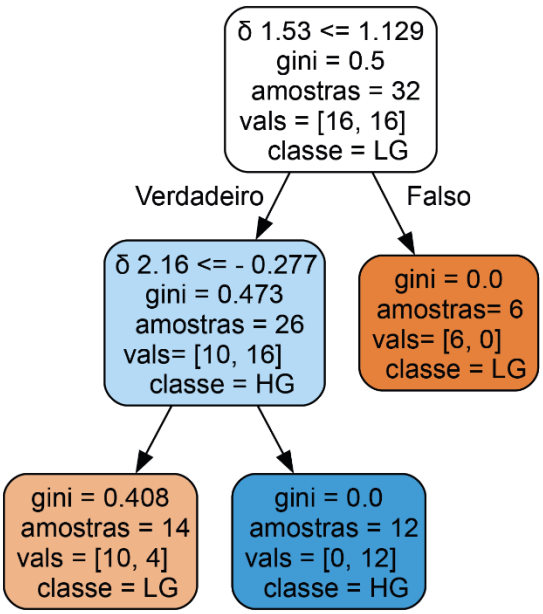
Na etapa de estadiamento, o conjunto de treinamento, agora com 32 amostras e balanceado entre as duas classes, LG e HG, teve sua dimensionalidade reduzida a partir do algoritmo SFM-ETC, resultando em sete variáveis selecionadas (Figura 38).

Figura 38. Gráfico de caixas, mostrando a distribuição das variáveis selecionadas pelo algoritmo SFM-ETC para o Conjunto de dados de estadiamento.



O novo conjunto de dados com as sete variáveis, apresentadas no gráfico de caixas da Figura 38, foi utilizado para o treinamento do modelo de árvore de decisão. O melhor modelo de classificação possuía profundidade de dois e está apresentado na Figura 39.

Figura 39. Árvore de decisão da etapa de estadiamento. Cada nó apresenta a impureza de Gini, o número de amostras, a distribuição por classes: [LG, HG], a classe com a maioria das amostras.



Como pode ser visto na Figura 39, para o conjunto de treinamento, quatro amostras do grupo HG foram classificadas erroneamente como grupo LG. Além disso, apenas as variáveis δ 1,53 ppm e δ 2,16 ppm foram utilizadas para construir a árvore. O resultado para o conjunto de validação está apresentado na matriz de confusão apresentada no Quadro 5.

Quadro 5. Matriz de confusão para o modelo de árvore de decisão (DT, do inglês: *Decision Tree*) de estadiamento.

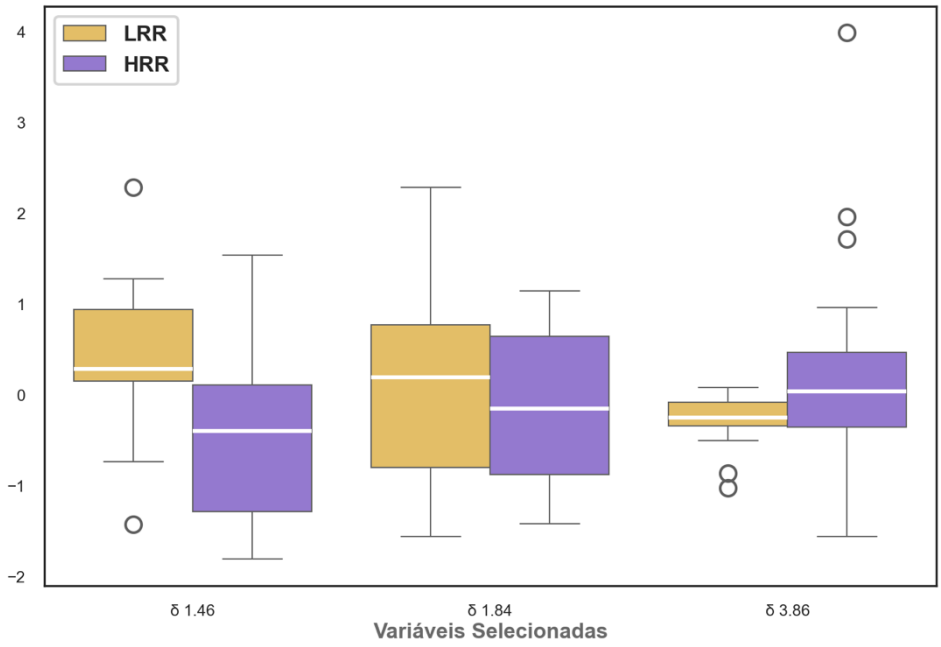
		Diagnóstico Clínico		
		LG	*HG*	*Total*
Modelo SFM-ETC-DT	*LG*	**16(5)**	4(1)	20(6)
	HG	0(0)	**12(3)**	12(3)
	Total	16(5)	16(4)	

A partir do Quadro 5, são verificados os valores de exatidão, sensibilidade e especificidade iguais a 88,9%, 75% e 100%, respectivamente, considerando o grupo HG como positivo. Além disso, o fator Kappa é de 0,769. O grupo HG mostrou-se mais disperso e, portanto, mais difícil de modelar, isso pode ser claramente observado pelo número de erros de classificação nesse grupo.

5.2.2 Risco de recidiva bioquímica

Na etapa de predição do risco de recidiva bioquímica, o conjunto de treinamento, também com 16 amostras em cada uma das classes, LRR e HRR, teve sua dimensionalidade reduzida a partir do algoritmo SFM-ETC. A SF selecionou três variáveis, onde suas distribuições estão dispostas no gráfico de caixas na Figura 40.

Figura 40. Gráfico de caixas, mostrando a distribuição das variáveis selecionadas pelo algoritmo SFM-ETC para o Conjunto de dados de risco de recidiva.



A nova matriz de dados, com três variáveis, foi utilizada para treinar o modelo kSVM, utilizando um *kernel* RBF (Raschka e Mirjalili, 2019) e o valor de C igual a $1 \cdot 10^{-5}$. O resultado do treinamento e validação do modelo SFM-ETC-SVM está apresentado na matriz de confusão, no Quadro 6.

Quadro 6. Matriz de confusão para o modelo de SVM de risco de recidiva bioquímica.

		Diagnóstico Clínico		
		LRR	*HRR*	*Total*
Modelo SFM-ETC-kSVM	*LRR*	**14(3)**	3(1)	17(4)
	HRR	2(1)	**13(4)**	15(5)
	Total	16(5)	16(4)	

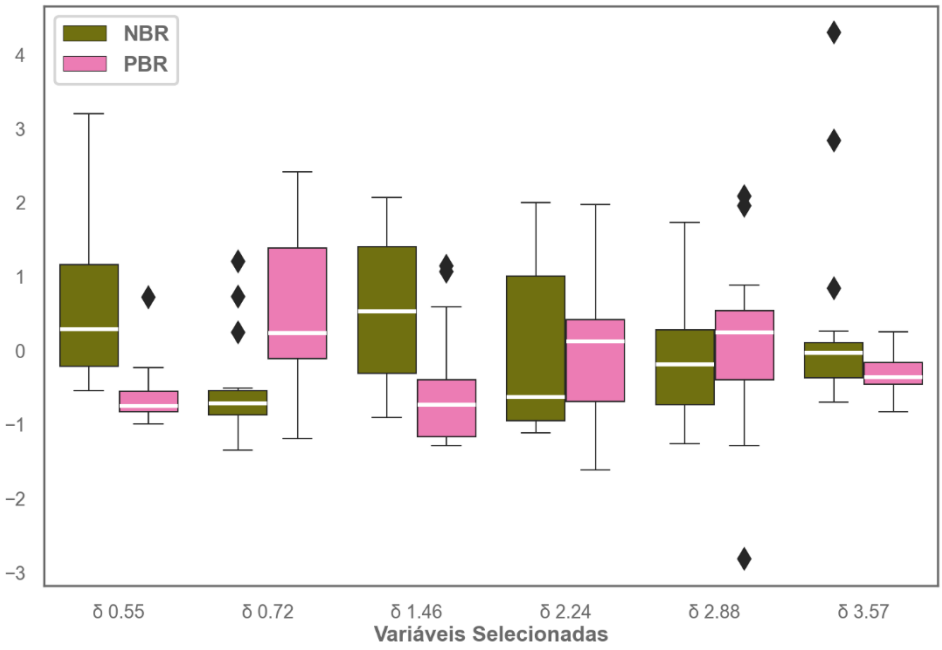
A partir dos resultados mostrados no Quadro 6, pode-se verificar que, considerando o grupo HRR como o grupo positivo, o modelo possuiu, na etapa de validação, exatidão de 77,8%, sensibilidade de 75% e especificidade de 80%.

Como nenhum dos modelos lineares, como LR e LDA, assim como o SVM utilizando um *kernel* linear foram capazes de modelar os dados desse conjunto de dados, pode se inferir que as duas classes não são linearmente separáveis. Porém, modelos não-paramétricos, ou modelos que utilizam um truque de *kernel*, que é o caso do kSVM, podem contornar esse tipo de problema (Raschka e Mirjalili, 2019). Porém, a maioria desses modelos não permite acessar diretamente a partir dos coeficientes, como é o caso do LDA, LR e do SVM, a importância das variáveis, podendo ser necessário a utilização de um outro algoritmo para auxiliar nessa tarefa (Hogan *et al.*, 2021).

5.2.3 Recidiva Bioquímica

Na última dessas etapas, como apresentado na Tabela 7, o conjunto de treinamento possuía 16 amostras de cada classe, (NBR e PBR), após o SMOTE e, assim como os anteriores, foi submetido a uma seleção de variáveis utilizando o método SFM-ETC. O método de redução da dimensionalidade selecionou seis variáveis e a distribuição destas está disposto no gráfico de caixas na Figura 41.

Figura 41. Gráfico de caixas, mostrando a distribuição das variáveis selecionadas pelo algoritmo SFM-ETC para o Conjunto de dados de recidiva bioquímica.



A matriz com dimensionalidade reduzida foi utilizada no treinamento do modelo LDA, para a qual, a matriz de confusão se encontra no Quadro 7.

Quadro 7. Matriz de confusão para o modelo de LDA de prognóstico de recidiva bioquímica.

		Diagnóstico Clínico		
		NBR	*PBR*	*Total*
Modelo SFM-ETC-LDA	*NBR*	**12(3)**	2(0)	14(3)
	PBR	4(2)	**14(5)**	18(7)
	Total	16(5)	16(5)	

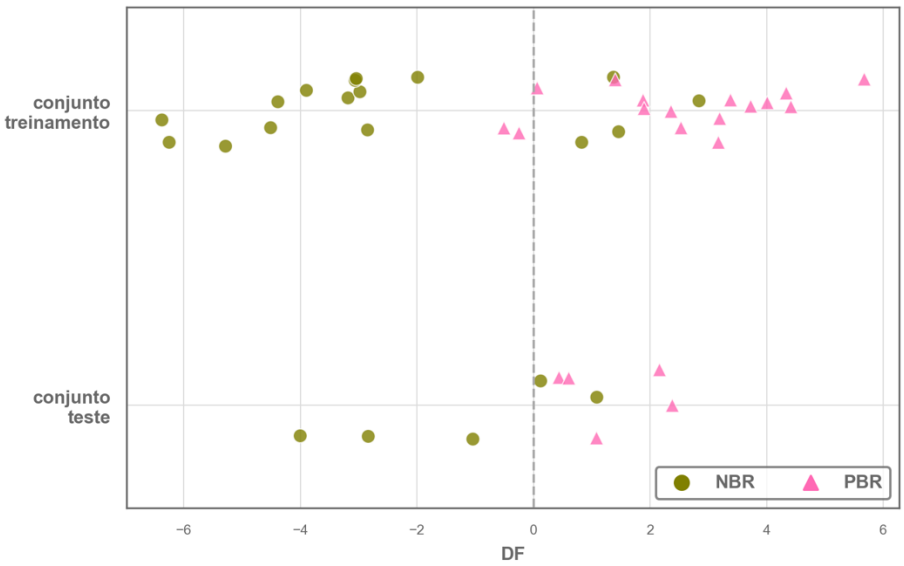
A partir dos resultados mostrados no Quadro 7, foi possível calcular as figuras de mérito exatidão, sensibilidade e especificidade, 80%, 100% e 60%, respectivamente, além do fator Kappa, igual a 0,60, para o modelo de prognóstico de recidiva bioquímica.

Na análise de PCA (Figura 37C) já era possível observar que a distribuição das amostras da classe NBR eram mais dispersas que da classe positiva, por isso o menor valor de especificidade, devido a dificuldade do algoritmo LDA em modelar esse grupo.

A DF da LDA é mostrada na Equação 42 e no gráfico de escores, na Figura 42.

$$DF = -0{,}978X_{0{,}55} + 0{,}484X_{0{,}72} - 1{,}818X_{1{,}46} + 0{,}989X_{2{,}24} + 0{,}893X_{2{,}88} - 2{,}001X_{3{,}57} \quad (42)$$

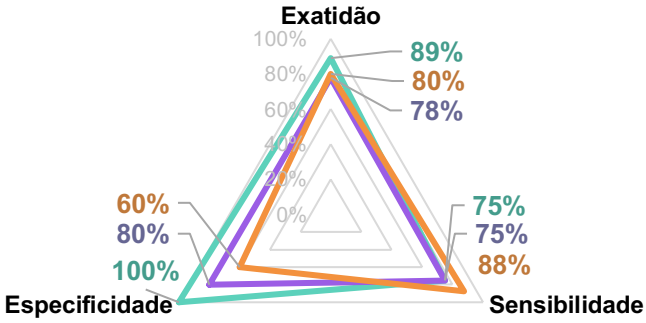
Figura 42. Função de decisão (DF) para os conjuntos de treinamento e teste do LDA para o prognóstico de recidiva bioquímica



Segundo a Equação 42, as variáveis mais importantes para a classificação são os deslocamentos químicos em δ 3,57 ppm e δ 1,46 ppm, que possuem valores maiores para a classe NBR, e em δ 2,24 ppm, que possui maiores valores para a classe PBR.

Os modelos de classificação com melhores figuras de mérito na predição, para os três conjuntos de dados – Estadiamento, Risco de Recidiva e Recidiva Bioquímica, foram obtidos após sobreamostragem e seleção de variáveis e, com exceção do LDA utilizado para classificar o conjunto de dados de Recidiva Bioquímica, foram realizados com um estimador que possuía algum tipo de regularização. As figuras de mérito das predições dos modelos de classificação para essas etapas se encontram na Figura 43.

Figura 43. Gráfico de radar com a exatidão, sensibilidade e especificidade das predições dos modelos de classificação DT para o conjunto de Estadiamento (verde), SVM para o conjunto de Risco de Recidiva (roxo) e LDA para o conjunto de Recidiva Bioquímica (laranja).



Por se tratar de um estudo preliminar, e conter um número de amostras reduzido, não foram realizadas maiores inferências sobre os metabólitos de maior importância para essa classificação.

A fim de avaliar se há melhora nas performances de predição de algoritmos de classificação, ao utilizar as técnicas de pré-processamento: sobreamostragem e seleção de variáveis, além de modelos regularizados, foram utilizados conjuntos de dados com poucas amostras e/ou desbalanceados previamente publicados. Os conjuntos utilizados para esse fim foram os conjuntos de dados para diagnóstico de asma em gatos (5.3) e o conjunto de dados acerca da imunização da varíola (5.4). Para tal, foi realizado o mesmo fluxo de trabalho quimiométrico mostrado na Figura 25 e realizado para o conjunto de dados de Diagnóstico de CaP.

5.3 DIAGNÓSTICO DE ASMA EM GATOS

O Conjunto de Dados de Diagnóstico Asma em Gatos[7] é proveniente do estudo intitulado *"Noninvasive Recognition and Biomarkers of Early Allergic Asthma in Cats"*, de Fulcher e colaboradores (2016), sediados na Universidade de Missouri-Columbia. O coorte do estudo consiste em amostras de 53 gatos domésticos, *Felis catus*, antes (Grupo 0) e depois de ter asma induzida pelo alérgeno da grama Bermuda (Grupo 1) (Fulcher *et al.*, 2016).

As amostras consistem em condensado de ar exalado pelos animais. As 106 amostras foram submetidas a análise por RMN de ^{1}H, usando um espectrômetro Bruker de 18 T, operando a 800 MHz no canal do ^{1}H. Para suprimir o sinal da água os pesquisadores utilizaram a sequência de pulso W5-Watergate (Furihata, Shimotakahara e Tashiro, 2008), além disso foi utilizada também a pressaturação da ressonância da água.

Após a remoção do sinal da água, entre 3,8 e 4,5 ppm, a matriz de dados final foi criada utilizando a resolução inteira do espectro. O conjunto final possui as dimensões 106 × 21379. Para esse estudo, o conjunto de dados foi desbalanceado com a retirada de quinze amostras do grupo 0.

Por conta do número muito grande de variáveis, seria inviável para um computador comum conseguir fazer as operações posteriores – seleção de variáveis e classificação. Por conta disso, um filtro que computa a estatística F na ANOVA foi utilizado para reduzir a

[7] O conjunto de dados para o estudo citado pode ser encontrado no Metabolomics Workbench pelo endereço web: https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000406&StudyType=NMR&ResultType=2

dimensionalidade, antes de os dados serem colocados no laço. O limite para manter as variáveis foi de 95% de confiança utilizando o p-valor do teste.

A pesquisa metabolômica publicada por Fulcher e colaboradores (2016) tinha como objetivo a construção de modelos de classificação por análise discriminante por quadrados mínimos parciais sem (PLS-DA) e com filtro ortogonal de sinal (OSC-PLS-DA), para discriminar o metabotipo de 53 gatos sem (grupo 0) e com asma recente induzida (grupo 1).

Os resultados publicados pelos pesquisadores estão apresentados na Tabela 8.

Tabela 8. Figuras de mérito dos modelos PLS-DA e OSC-PLS-DA para os dados de Fulcher e colaboradores em seu artigo sobre diagnóstico de asma em gatos: a sensibilidade e especificidade para o conjunto de teste.

	PLS-DA	**OSC-PLS-DA**
Sensibilidade.....	94,1%	94,1%
Especificidade.....	94,1%	100%
Q^2.....	0,698	0,844

Fonte dos resultados: (Fulcher *et al.*, 2016)

A intenção do escopo desse trabalho não é comparar os resultados obtidos pelos autores originais e os resultados apresentados na seguinte pesquisa, uma vez que o conjunto de dados foi alterado para possuir algum tipo de desbalanceamento de classes. Para isso, após a importação dos dados, foram eliminadas aleatoriamente 15 amostras do grupo 0, e a partir desse desbalanceamento entre classes, avaliar se há diferença nos parâmetros dos modelos de classificação com e sem reamostragem.

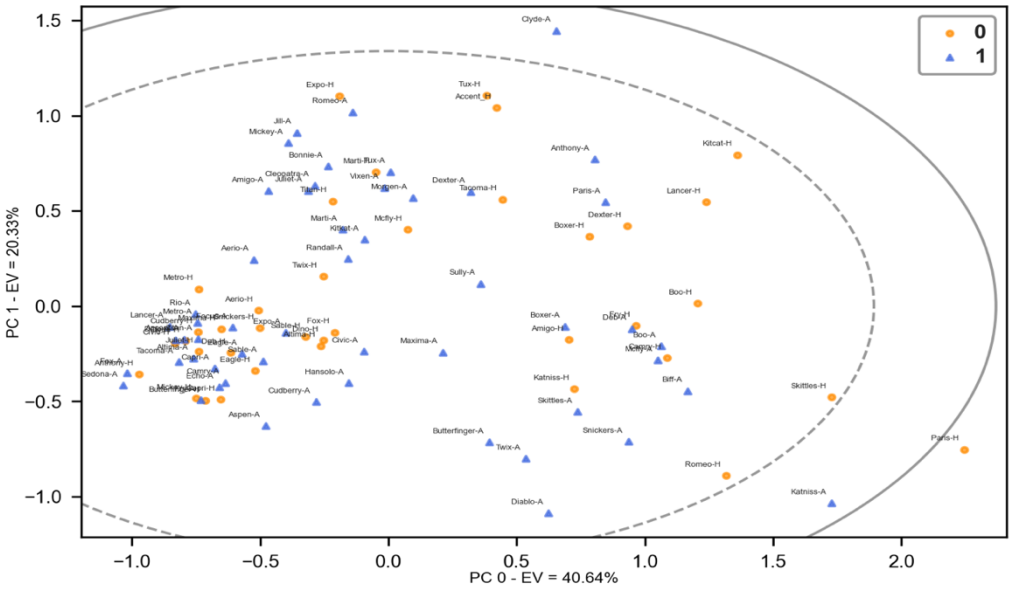
Um modelo OSC-PLS-DA ou Análise Discriminante por Projeções Ortorgonais para Estruturas Latentes (OPLS-DA, do inglês: *Orthogonal Projections to Latent Structures Discriminant Analysis*) (Trygg e Wold, 2002) foi treinado com os dados após a remoção das 15 amostras e ilustra a diferença causada pelo desbalanceamento das classes. Os resultados são exibidos na Tabela 9.

Tabela 9. Figuras de mérito dos modelos PLS-DA e OSC-PLS-DA para os dados de Fulcher e colaboradores, após a remoção de 15 amostras do grupo 0: a sensibilidade e especificidade para o conjunto de teste.

	OSC-PLS-DA
Sensibilidade.....	100%
Especificidade.....	8,3%
Q^2.....	0,266

Na Tabela 9 pode ser visualizada a redução, principalmente na especificidade, devido ao enviesamento do modelo em relação ao grupo majoritário (Grupo 1), demonstrando a dificuldade do algoritmo em lidar com classes desbalanceadas. Os resultados completos para esse modelo estão apresentados no Apêndice E. Posteriormente, foi realizada uma EDA, para detecção de agrupamentos e amostras anômalas (Figura 44). Para mais informações sobre os modelos PLS-DA e OPLS-DA ver Apêndice D.

Figura 44. Escores da PCA do CONJUNTO DE DADOS PARA DIAGNÓSTICO DE ASMA EM GATOS. Adicionalmente são mostradas as elipses de T^2 de Hotelling de 95% e 99% de confiança.



A partir dos escores da PCA, apresentadas na Figura 44, vemos a presença de uma amostra fora da elipse de 99% de confiança utilizando a distância T^2 de Hotelling, Paris_H, que pertence à classe 0. Nesse caso, essa amostra foi considerada anômala e retirada das próximas etapas de análise. A matriz resultante foi separada em conjunto de treinamento e teste.

Após isso, a matriz passou por uma primeira etapa de seleção de variáveis, utilizando a estatística F da ANOVA, no conjunto de treinamento, que continha 65% do número total de amostras (os 35% restantes foram utilizados para a geração do conjunto de teste). Como o teste é feito consecutivamente a cada variável, as variáveis que não apresentam diferenças estatística entre as médias dos dois grupos, no nível de confiança de 95%, são retiradas. Essa primeira redução de dimensionalidade foi utilizada devido ao grande tamanho da matriz, que possui 21379 deslocamentos químicos e não poderia ser analisada posteriormente, devido ao acúmulo de dados na memória do computador. Após a seleção de variáveis, a matriz passou a ter 5224 variáveis.

O conjunto já reduzido, antes de ser colocado no laço foi testado com os estimadores antes da sobreamostragem (Tabela 10). Utilizando o algoritmo SMOTE, os dois grupos foram balanceados e foi realizada uma nova etapa de EDA, para analisar as amostras sintéticas (Figura 45).

Figura 45. EDA dos dados após SMOTE para o Conjunto de dados para diagnóstico de asma em gatos. A) Os espectros em verde são das amostras geradas por SMOTE. B) Escores da PCA destacando as amostras sintéticas.

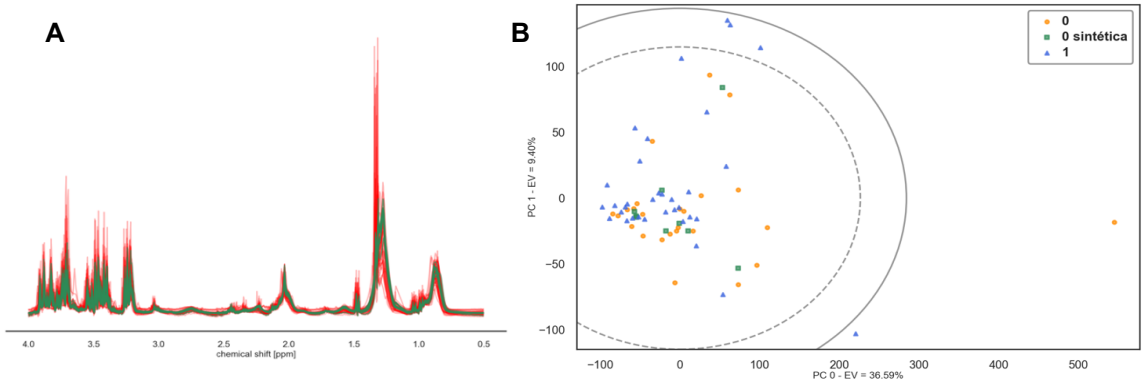


Tabela 10. Resultados para os classificadores do Conjunto de dados para diagnóstico de asma em gatos antes e depois do SMOTE. Valores entre parênteses, sem SF. Em negrito, destacam-se os algoritmos com melhor fator Kappa

Estimador	*Exatidão*	*Sensibilidade*	*Especificidade*	*AUROC*	*Kappa*
SVM	**0,84(0,84)**	**0,90(0,74)**	**0,77(1,00)**	**0,83(0,87)**	**0,67(0,70)**
LDA	**0,62(0,91)**	**1,00(1,00)**	**0,08(0,77)**	**0,54(0,88)**	**0,09(0,80)**
LR	**0,81(0,69)**	**0,84(1,00)**	**0,77(0,23)**	**0,81(0,62)**	**0,61(0,26)**
XGB	0,41(0,66)	0,00(0,42)	1,00(1,00)	0,50(0,71)	0,00(0,37)
KNN	0,50(0,72)	0,84(0,95)	0,00(0,38)	0,42(0,67)	0,00(0,36)

Assim como foi observado no Conjunto de Dados de Diagnóstico (vide item 5.1), as amostras sintéticas geradas pelo SMOTE para o Conjunto de dados para diagnóstico de asma em gatos também tendem a se agrupar próximo ao centro da distribuição dos dados (Figura 45B) e as amostras sintéticas possuem perfil semelhante a um espectro de RMN de ^{1}H original (Figura 45A).

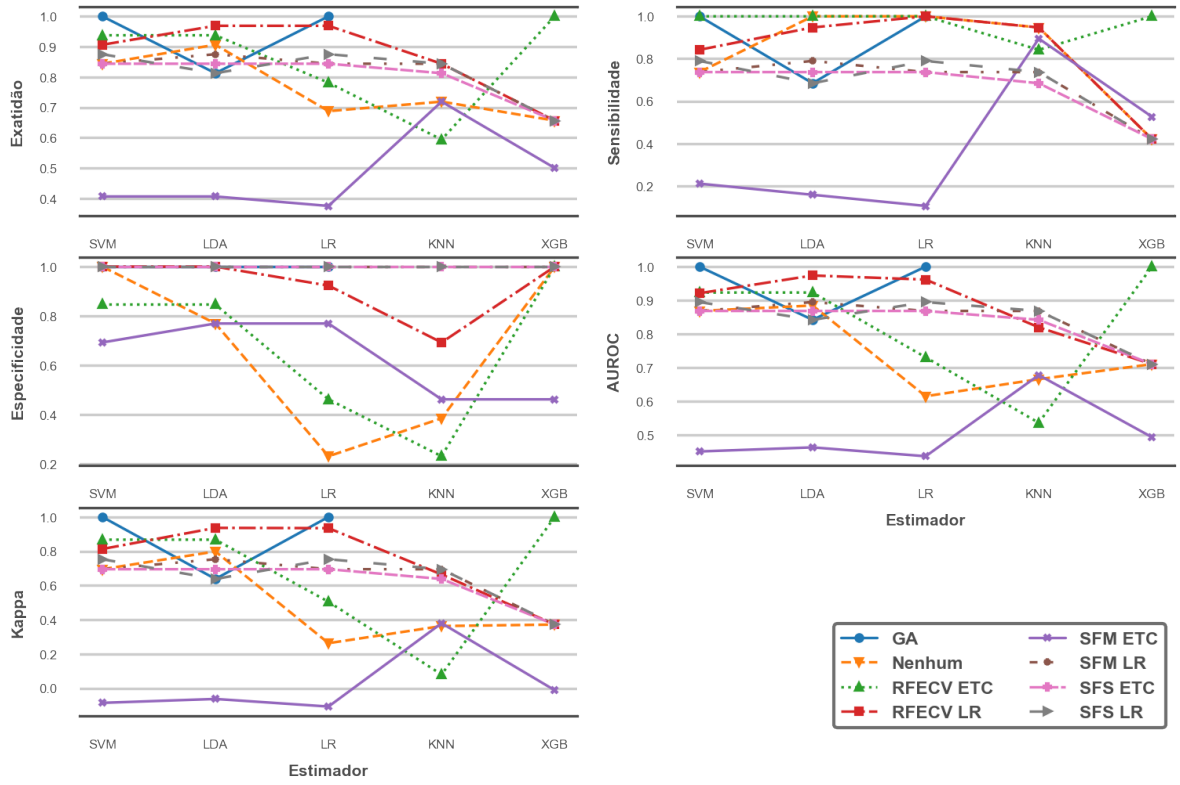
O conjunto de dados sobreamostrado foi então carregado no laço, combinando as diversas seleção de variáveis e estimadores. Os resultados apresentados para esse conjunto de dados se encontram na Figura 46. Os dez formalismos com maiores valores de fator Kappa se apresentam na Tabela 11 e todos os resultados podem ser conferidos no Apêndice B.

Tabela 11. Os dez formalismos com maiores valores de Kappa de Cohen após laço contendo os formalismos com SELEÇÃO DE VARIÁVEIS e classificadores para o CONJUNTO DE DADOS PARA DIAGNÓSTICO DE ASMA EM GATOS. Os valores em negrito destacam os modelos plausíveis.

Seletor de Variáveis	*Estimador*	*Exatidão*	*Sensibilidade*	*Especificidade*	*AUROC*	*Kappa*	*Nº de variáveis*
Nenhuma	LDA	0,91	1,00	0,77	0,88	0,80	NA
GA	SVM	1,00	1,00	1,00	1,00	1,00	1
	LR	1,00	1,00	1,00	1,00	1,00	2
RFECV-ETC	**XGB**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1565**
	SVM	0,94	1,00	0,85	0,92	0,87	1565
	LDA	0,94	1,00	0,85	0,92	0,87	1565
RFECV-LR	LDA	0,97	0,95	1,00	0,97	0,94	617
	LR	**0,97**	**1,00**	**0,92**	**0,96**	**0,93**	**617**
	SVM	0,91	0,84	1,00	0,92	0,81	617
SFM-LR	LDA	0,88	0,79	1,00	0,89	0,75	50

SFS-LR	SVM	0,88	0,79	1,00	0,89	0,75	24

Figura 46. Resultados do laço contendo os formalismos com SELEÇÃO DE VARIÁVEIS e classificadores após SMOTE para o CONJUNTO DE DADOS PARA DIAGNÓSTICO DE ASMA EM GATOS.



Como pôde ser observado na Figura 46 e nas tabelas 10 e 11, as três melhores combinações tiveram 100% de exatidão, GA-LR, GA-SVM e RFECV-ETC-XGB. Porém, o GA-LR e o GA-SVM resultaram num número muito pequeno de variáveis, duas e uma variável, respectivamente, tornando os modelos com pouca robustez.

O algoritmo XGB é de grande escalabilidade, sua dependência com o grande número de hiperparâmetros que ele possui pode dificultar a modelagem, porém, se bem realizada, as regularizações permitem um bom controle da variância, assim como a combinação de árvores de decisão reduzem o viés (Chen e Guestrin, 2016a; CHEN e GUESTRIN, 2017).

Porém, para uma análise metabolômica, a considerar que o XGB é um modelo que não fornece a importância das variáveis diretamente, a escolha do melhor formalismo ficaria para o REFCV-LR-LDA ou REFCV-LR-LR, que tiveram a performance, como mostrado na Tabela 9, próximas aos anteriores e a importância das variáveis pode ser extraída a partir dos coeficientes do modelo. Como já citado, as variáveis de interesse são de grande importância no entendimento da doença, uma vez que, podemos analisar os metabólitos que possuem maior influência na discriminação dos grupos.

A seleção de variáveis realizada para o LDA resultou num número de variáveis bem superior ao número de amostras, como mostrado na Tabela 11. O cálculo de LDA a partir da sua matriz de covariância e a inversão dessa matriz não modela bem esse tipo de dados (Ferreira, 2015; Pedregosa, F. *et al.*, 2011). Outras discussões sobre os formalismos e seus resultados serão realizadas conjuntamente após o item 5.4.3.

5.4 CONJUNTO DE DADOS DE AVALIAÇÃO DA IMUNIZAÇÃO DA VARÍOLA

Do estudo *"Metabotypes of Subjects with Adverse Reactions Following Vaccination"*, de McClenathan e colaboradores (2017), da Universidade da Carolina do Norte, foi obtido Conjunto de dados acerca da imunização da varíola[8]. O estudo tinha como objetivo avaliar as alterações metabólicas entre os diversos efeitos adversos e pessoas que não tiveram nenhum evento negativo após a imunização de varíola. Os indivíduos foram categorizados em quatro grupos: cinco pacientes que apresentaram miocardite após a vacinação, trinta pacientes que apresentaram elevação da troponina, que é classificada como miocardite subclínica, 31 pacientes que apresentaram outros sintomas sistêmicos, como febre, artralgias, febres, mialgias e/ou dores de cabeça, após a imunização, além de um grupo com 34 indivíduos que não reportaram sintomas após a imunização. As amostras dos pacientes foram coletadas em dois momentos, antes e após a imunização. Para continuar com dados binários, combinamos todos os indivíduos que apresentaram algum tipo de sintoma, 66, como grupo 1, e os 34 que não apresentaram, como grupo 0, apenas com os dados após a imunização (McClenathan *et al.*, 2017).

O biofluido utilizado no estudo foi o soro sanguíneo. As amostras foram analisadas utilizando um espectrômetro Bruker de 16 T (700 MHz para o ^{1}H). A sequência de pulsos utilizada foi 1D NOESY com pressaturação do sinal da água e as medidas foram realizadas a 298 K.

Após a exclusão das regiões da água, entre 4,66 e 5,16 ppm, e metanol, entre 3,30 e 3,37 ppm, que foi utilizado no preparo da amostra. Os espectros passaram pelo processo de *bucketing* com um *bin* de 0.04 ppm de largura entre 0,07 ppm e 8,50 ppm. Ao fim do pré-tratamento dos espectros, a matriz resultante tinha dimensão 100 × 183.

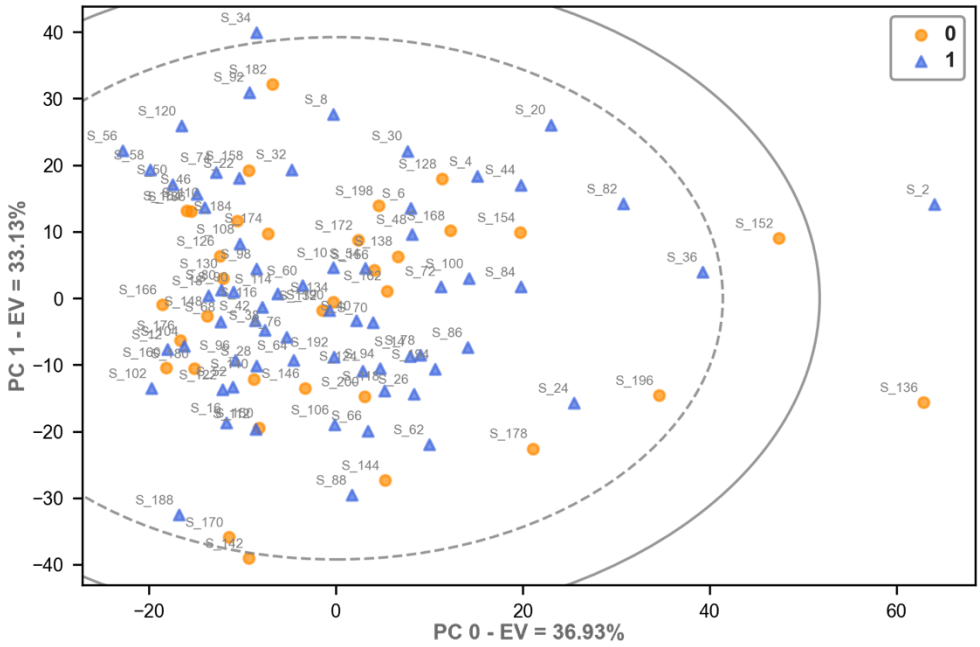
No artigo de McClenathan e colaboradores (2017), objetivando a descoberta de novos biomarcadores e o metabotipo de reações adversas à imunização de varíola, foram construídos modelos PLS-DA comparando os metabólitos no soro sanguíneo antes e depois da imunização.

[8] O conjunto de dados para o estudo citado pode ser encontrado no Metabolomics Workbench pelo endereço web: https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000339

Como citado anteriormente, para testar os formalismos propostos nesse trabalho, foram utilizados apenas os dados pós imunização, com os 66 pacientes que tiveram algum tipo de reação no grupo 1 e os 34 que não tiveram no grupo 0. Uma abordagem muito distinta da proposta pelo autor do trabalho, por isso não serão trazidos os resultados obtidos pelo autor, para mais informações ler o artigo original (McClenathan *et al.*, 2017).

Primeiramente, as amostras foram autoescaladas e passaram por SNV e finalmente foi realizada uma PCA (Figura 47), para visualizar melhor os dados e verificar a presença de amostras anômalas.

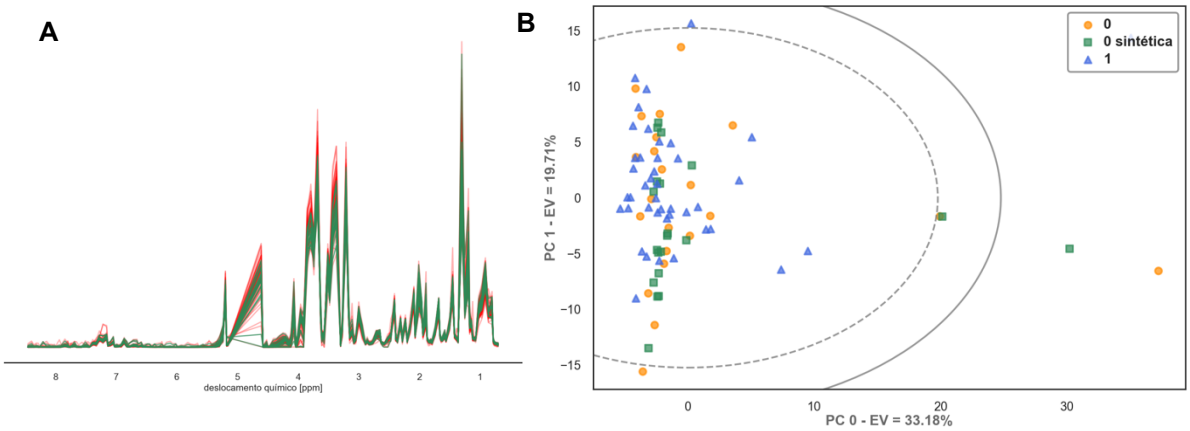
Figura 47. Escores da PCA do Conjunto de dados acerca da imunização da varíola. Adicionalmente são mostradas as elipses de T^2 de Hotelling de 95% e 99% de confiança.



Nos escores da PCA (Figura 47), fica clara a presença de duas amostras fora da elipse de 99% de confiança de T2 de Hotelling – amostras S_2 e S_136, que foram removidas do conjunto de dados. Com as amostras restantes, foi realizada a separação dos conjuntos de treinamento e de teste. O conjunto de treinamento possuía 42 amostras do grupo 1 e 21 amostras do grupo 0. Já o conjunto de teste possuía 23 amostras do grupo 1 e 12 amostras do grupo 0.

O conjunto de treinamento, que contém 65% do número total de amostras, foi utilizado, inicialmente, para treinar todos os modelos de classificação antes da sobreamostragem e de qualquer seleção de variáveis. Utilizando o algoritmo SMOTE, os dois grupos foram balanceados e foi realizada uma nova etapa de EDA, para analisar as amostras sintéticas (Figura 48).

Figura 48. EDA dos dados após SMOTE para o Conjunto de dados acerca da imunização da varíola. A) Os espectros em verde são das amostras geradas por SMOTE. B) Escores da PCA destacando as amostras sintéticas.

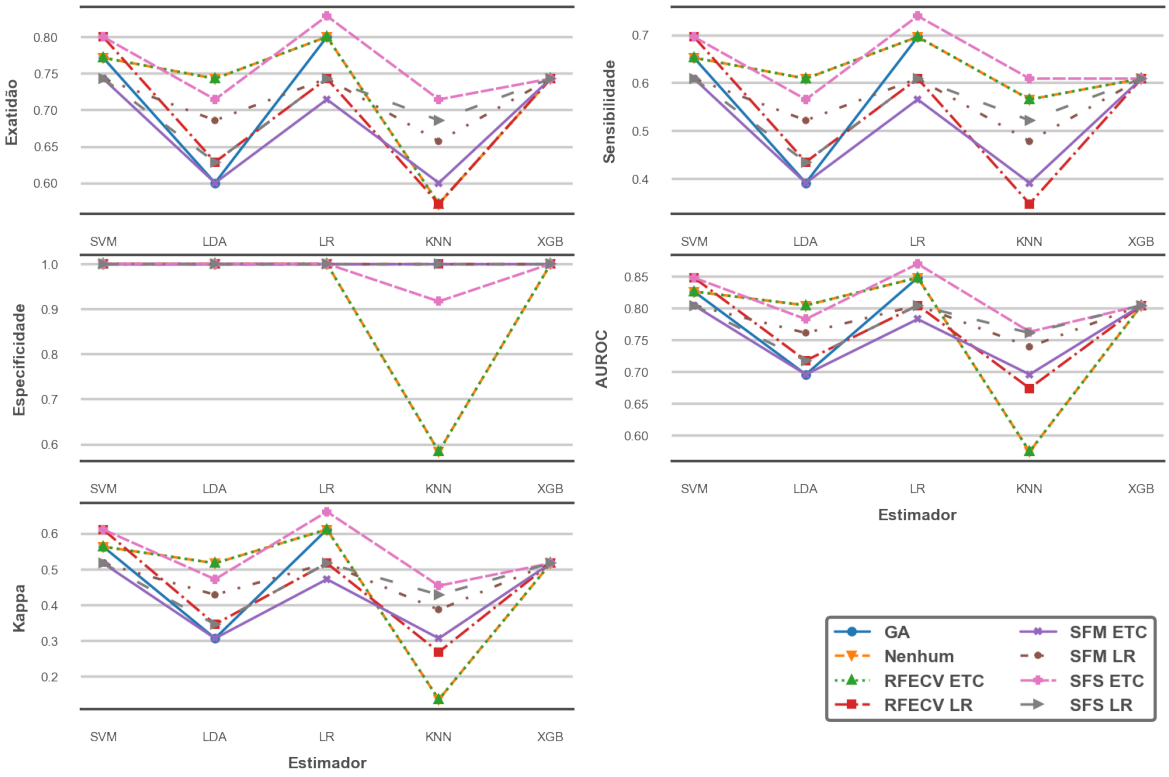


Foram geradas 21 amostras sintéticas da classe 0 no conjunto de treinamento, igualando ao número de amostras originais do grupo 1, 42 amostras. Na Figura 48A, vê-se que o SMOTE não utilizou de artifício matemático que criasse amostras que não fossem compatíveis com as demais amostras, ou seja, com as mesmas características de espectro das demais. Na Figura 48B, é notável que o conjunto de treinamento possui amostras muito dispersas em relação ao centro da distribuição, como o SMOTE é um algoritmo que utiliza do método de vizinhos mais próximos, ele acabou gerando novas amostras nessa região mais dispersa também. Em seguida, os dados de treinamento e teste sobreamostradas foram aplicados ao laço. As figuras de mérito, na etapa de validação externa, tanto dos modelos antes e depois da sobreamostragem sem seleção de variáveis, quanto do laço com todas as possibilidades estão dispostas na Tabela 12 e na Figura 49.

Tabela 12. Resultados para os classificadores do CONJUNTO DE DADOS ACERCA DA IMUNIZAÇÃO DA VARÍOLA antes e depois do SMOTE. Valores entre parênteses, sem SF. Em negrito destacam-se os algoritmos com melhor fator Kappa.

Estimador	***Exatidão***	***Sensibilidade***	***Especificidade***	***AUROC***	***Kappa***
SVM	0,66(0,77)	1,00(0,65)	0,00(1,00)	0,50(0,82)	0,00(0,56)
LDA	0,69(0,74)	0,96(0,61)	0,17(1,00)	0,56(0,80)	0,15(0,52)
LR	0,66(0,80)	1,00(0,70)	0,00(1,00)	0,50(0,85)	0,00(0,61)
XGB	**0,77(0,74)**	**0,91(0,61)**	**0,50(1,00)**	**0,71(0,80)**	**0,45(0,52)**
KNN	0,71(0,57)	0,91(0,56)	0,33(0,58)	0,62(0,57)	0,28(0,13)

Figura 49. Resultados do laço contendo os formalismos com SELEÇÃO DE VARIÁVEIS e classificadores após SMOTE para o Conjunto de dados acerca da imunização da varíola.



Na Tabela 12, fica evidenciada a dificuldade que os modelos apresentavam antes da realização do SMOTE em modelar o grupo 0. Isso ocorre devido ao viés gerado pelo número maior de amostras da classe 1 no conjunto de dados (Wu e Fang, 2020). Esse problema é invertido quando há a reamostragem.

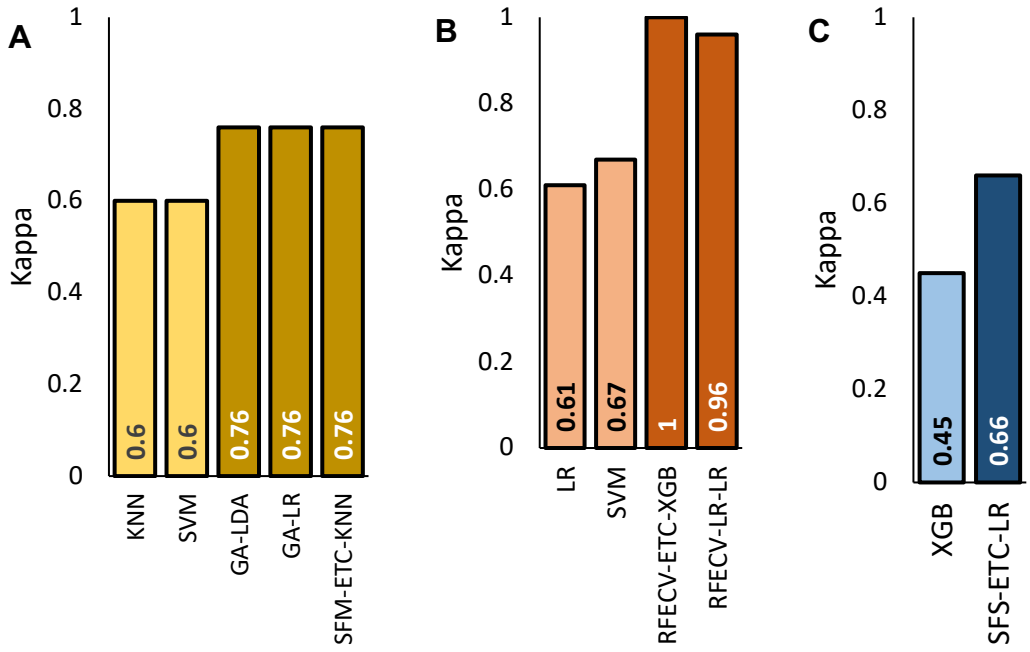
Na Figura 49, pode ser observado o quanto a maioria das combinações de formalismos consegue modelar bem a classe 0. Mesmo com a geração de amostras da classe 0, que se mostraram dispersas na PCA, esse grupo se mostrou mais fácil de ressaltar nesses formalismos. Isso pode ser explicado pela origem das amostras, uma vez que as amostras da classe 0 partiram de um grupo único, enquanto as amostras da classe 1 foram obtidas a partir da reunião de três grupos com condições distintas entre si, tornando uma modelagem complexa para a classificação entre os dois grupos.

O subconjunto dos dados, gerados após a SFS-ETC foi a que obteve os maiores valores de fator Kappa e de AUROC em quase todos os modelos, com exceção do LDA, Inclusive, a melhor combinação foi a SFS-ETC-LR, com exatidão de 0,83, sensibilidade de 0,74, especificidade de 1,00, fator Kappa de 0,66 e AUROC de 0,87. Todos os resultados podem ser consultados no Apêndice C.

Porém, para esses estudos, com os Conjunto de Dados de Estadiamento, Conjunto de Dados de Risco de Recidiva e Conjunto de Dados de Recidiva Bioquímica, além do Conjunto de Dados de Diagnóstico de CaP, fica claro que nos espectros de RMN do soro há informação sobre a alteração do metabolismo causado pelo CaP, além da sua gradação. Portanto, esses achados demonstram que a metabonômica baseada em espectros de RMN de ^{1}H do soro de sangue pode ser uma ferramenta no auxílio ao diagnóstico, estadiamento e prognóstico de recidiva bioquímica do CaP.

Para os seis conjuntos de dados que foram abordados, os modelos de classificação com melhores performances para o conjunto de validação foram treinados após a etapa de reamostragem e seleção de variáveis. Além disso, os classificadores LR, LDA, SVM e XGB que foram otimizados possuem ao menos um método de regularização ativo na modelagem. Em destaque, na Figura 50, encontram-se os valores de Kappa de Cohen dos melhores modelos de classificação com SMOTE e seleção de variáveis e sem essas etapas, para os conjuntos de dados de Diagnóstico de CaP, Diagnóstico de Asma em Gatos e Avaliação da Imunização da Varíola.

Figura 50. Comparação do valor de Fator Kappa de Cohen dos melhores modelos de classificação sem os pré-processamentos com SMOTE e seleção de variáveis (cor mais clara) e com SMOTE e seleção de variáveis (cor mais escura) para: A) conjunto de dados de Diagnóstico de CaP, B) Diagnóstico de Asma em Gatos e C) Avaliação da Imunização da Varíola.



Essas evidências, citadas no parágrafo anterior, mostram que, para dados metabonômicos, que muitas vezes possuem poucas amostras, e vindo de dados espectrais de alta dimensionalidade, como os espectros RMN de ^{1}H, exemplificados nos seis conjuntos de dados utilizados nesse estudo, a seleção de variáveis pode ser de grande valia na redução do

efeito de sobreajuste, que esse tipo de dado pode acometer os modelos de classificação (Ko, Choi e Ahn, 2021).

Além disso, conjuntos de dados, como no Conjunto de Dados de Avaliação da Imunização da Varíola, que possuem um grande desbalanceamento entre as classes (a classe 0 possuía aproximadamente metade do número de amostras da classe 1), tendem a apresentar uma forte tendência na modelagem para o grupo majoritário, isso aumenta a variância do modelo, reduzindo sua capacidade de predição de novas amostras (Rodrigues, Luna e Pinto, 2023). Para essa finalidade, o SMOTE, como método de sobreamostragem, se mostrou uma poderosa alternativa na redução do sobreajuste.

Ainda sobre os modelos treinados, a regularização também foi um aspecto chave na diminuição da variância dos modelos, penalizando modelos mais complexos, como no caso do XGB, ou reduzindo a complexidade a partir da penalização de variáveis com alto peso (Ledoit e Wolf, 2004; Raschka e Mirjalili, 2019; Zhao *et al.*, 2020).

Em geral, esses aspectos mostram que há uma grande variedade de possibilidades no que diz respeito à aprendizagem de máquina e sua utilização para fins da área de saúde, como na metabonômica baseada em RMN de ^{1}H.

6 CONCLUSÃO

O presente trabalho tinha como objetivos demonstrar a capacidade da seleção de variáveis, modelos regularizados e do SMOTE em reduzir o efeito de sobreajuste causado por matrizes de grande dimensionalidade e poucos exemplos, utilizando (1 e 2) dois conjuntos de dados de metabonômica baseada em RMN de ^{1}H. Além disso, utilizar essas técnicas para criar um modelo de classificação que pudesse discriminar entre (3) homens portadores ou não portadores de CaP, (4) estadiar pacientes acometidos com CaP, (5) discriminar pacientes a partir do risco de recidiva bioquímica e (6) prognosticar a recidiva bioquímica do CaP.

Nos seis casos citados, os modelos de classificação que foram treinados após a seleção de variáveis, SMOTE e regularização possuem melhor desempenho na predição de dados não vistos no treinamento do que os modelos treinados sem passar por esse tipo de pré-processamento, reamostragem e redução de dimensionalidade, e escolha adequada do modelo de classificação a partir de algoritmos com penalização.

Acerca da matriz de dados que compõe o problema de diagnóstico do CaP, após uma busca exaustiva pelo melhor formalismo, o GA-LDA utilizando uma regularização por encolhimento, apresentou o melhor desempenho dentre as trinta e oito combinações buscadas. O GA-LDA apresentou sensibilidade e especificidade de 92% e 83%, respectivamente, e exatidão de 88%.

Acerca dos conjuntos de dados composto apenas por pacientes portadores de CaP, o modelo de estadiamento, árvore de decisão após o SFM-ETC, possuiu exatidão, sensibilidade e especificidade iguais a 88,9%, 75% e 100%. Além disso, o modelo de risco de recidiva bioquímica, SVM após o SFM-ETC, possuiu exatidão, de 77,8%, sensibilidade de 75% e especificidade de 80%. E finalmente, para o modelo de prognóstico de recidiva bioquímica, LDA após o SFM-ETC, possuiu exatidão, sensibilidade e especificidade, 80%, 87,5% e 60%.

Para o diagnóstico de CaP e os estudos preliminares em estadiamento, risco de recidiva bioquímica e prognóstico de recidiva bioquímica do CaP, foram encontradas evidências que as informações sobre essa doença estão presentes nos espectros de RMN de ^{1}H do soro.

Além disso, na etapa de diagnósito de CaP, a partir dos metabólitos identificados, nota-se uma perturbação nas quantidades de lipídios e carboidratos causada pela alta demanda energética e forte influência do ciclo de neoglicogênese. Ademais, há também perturbação nos níveis de alguns aminoácidos, que são consumidos como blocos construtores no crescimento acelerado das células tumorais, e nos níveis de lactato, que participam da glicólise aeróbia, conhecido como efeito Warburg.

REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, F.; ADRIAENS, M. The metamorphosis of analytical chemistry. **Analytical and Bioanalytical Chemistry**, v. 412, n. 15, p. 3525–3537, 2020.

ANTCLIFFE, D.; GORDON, A. C. Metabonomics and intensive care. **Critical Care**, v. 20, n. 1, p. 1–7, 2016.

ARAUJO, L. C. N. **Estudo Piloto: Diagnóstico do câncer urológico por via metabonômica**. [s.l.] Universidade Federal de Pernambuco, 2016.

ATTA-UR-RAHMAN; CHOUDHARY, M. I.; ATIA-TUL-WAHAB. The Basics of Modern NMR Spectroscopy. *Em*: **Solving Problems with NMR Spectroscopy**. [s.l.] Elsevier, 2016. p. 1–34.

AWAD, M.; FRAIHAT, S. Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. **Journal of Sensor and Actuator Networks**, v. 12, n. 5, 2023.

BA, F. *et al.* Classification and Identification of Contaminants in Recyclable Containers Based on a Recursive Feature Elimination-Light Gradient Boosting Machine Algorithm Using an Electronic Nose. **Micromachines**, v. 14, n. 11, p. 2047, 2023.

BARROS, C. J. P. **Metabonômica Baseada em RMN como Ferramenta para Discriminação de Grãos de Soja Irradiados & Diagnóstico de Hepatites e Fibrose Hepática**. [s.l.] Universidade Federal de Pernambuco, 2017.

BEINECKE, J.; HEIDER, D. Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. **BioData Mining**, v. 14, n. 1, p. 1–11, 2021.

BISHT, B. *et al.* The potential of nuclear magnetic resonance (NMR) in metabolomics and lipidomics of microalgae- a review. **Archives of Biochemistry and Biophysics**, v. 710, p. 108987, out. 2021.

BJERRUM, J. T. VEITEN. **Metabonomics**. New York, NY: Springer New York, 2015. v. 1277

BRERETON, R. G. *et al.* Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. **Analytical and Bioanalytical Chemistry**, v. 409, n. 25, p. 5891–5899, 2017.

BUENO DE PAIVA, L. *et al.* Effects of RhoA and RhoC upon the sensitivity of prostate cancer cells to glutamine deprivation. **Small GTPases**, v. 12, n. 1, p. 20–26, 2 jan. 2021.

BYLESJÖ, M. *et al.* OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. **Journal of Chemometrics**, v. 20, n. 8–10, p. 341–351, 2006.

CALZOLARI, M. **sklearn-genetic**, jan. 2022.

CARLSSON, S. V.; VICKERS, A. J. Screening for Prostate Cancer. **Medical Clinics of North America**, v. 104, n. 6, p. 1051–1062, 2020.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, n. Sept. 28, p. 321–357, 1 jun. 2002.

CHEN, T.; GUESTRIN, C. **XGBoost: A scalable tree boosting system**Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...New York, NY, USA: ACM, 13 ago. 2016a

___. **XGBoost**Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...New York, NY, USA: ACM, 13 ago. 2016b

CHEN, T.; GUESTRIN, C. **XGboost**. Disponível em: <https://xgboost.readthedocs.io/en/stable/python/python_api.html>. Acesso em: 9 jan. 2024.

CHEN, W. *et al.* Metabonomic characteristics and biomarker research of human lung cancer tissues by HR1H NMR spectroscopy. **Cancer Biomarkers**, v. 16, n. 4, 2016.

CHO, W. J. *et al.* Clinicopathological implications of histological mapping in radical prostatectomy specimens. **Pathology - Research and Practice**, v. 243, p. 154334, mar. 2023.

CHONG, J.; WISHART, D. S.; XIA, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. **Current Protocols in Bioinformatics**, v. 68, n. 1, p. 1–128, 2019.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 2 abr. 1960.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

DANIEL, J.; MARTIN, J. H. Logistic Regression. *Em*: **Speech and Language Processing**. 3rd. ed. [s.l: s.n.]. p. 25.

DANIELS, A.; WILLIAMS, R. J. P.; WRIGHT, P. E. Nuclear magnetic resonance studies of the adrenal gland and some other organs. **Nature**, v. 261, n. 5558, p. 321–323, 1 maio 1976.

DELAFIORI, J. *et al.* Covid-19 Automated Diagnosis and Risk Assessment through Metabolomics and Machine Learning. **Analytical Chemistry**, v. 93, n. 4, p. 2471–2479, 2021.

DIAMAND, R. *et al.* Risk stratification for early biochemical recurrence of prostate cancer in the era of multiparametric magnetic resonance imagining-targeted biopsy. **The Prostate**, v. 83, n. 6, p. 572–579, 27 maio 2023.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification (2nd Edition)**. 2. ed. [s.l.] Wiley-Interscience, 2000.

EMWAS, A. H. *et al.* Recommended strategies for spectral processing and post-processing of 1D 1 H-NMR data of biofluids with a particular focus on urine. **Metabolomics**, v. 14, n. 3, 2018.

FERREIRA, M. M. C. *et al.* Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v. 22, n. 5, p. 724–731, set. 1999.

FERREIRA, M. M. C. **Quimiometria: Conceitos, Métodos e Aplicações**. 1. ed. Campinas - SP: Editora Unicamp, 2015.

FULCHER, Y. G. *et al.* Noninvasive Recognition and Biomarkers of Early Allergic Asthma in Cats Using Multivariate Statistical Analysis of NMR Spectra of Exhaled Breath Condensate. **PLOS ONE**, v. 11, n. 10, p. e0164394, 20 out. 2016.

FURIHATA, K.; SHIMOTAKAHARA, S.; TASHIRO, M. An efficient use of the WATERGATE W5 sequence for observing a ligand binding with a protein receptor. **Magnetic Resonance in Chemistry**, v. 46, n. 9, p. 799–802, 2008.

FURMAŃCZYK, K. *et al.* Classification and feature selection methods based on fitting logistic regression to PU data. **Journal of Computational Science**, v. 72, n. October 2022, p. 102095, 2023.

GALVÃO, R. K. H.; ARAÚJO, M. C. U. DE; SOARES, S. F. C. Linear Regression Modeling: Variable Selection. *Em*: **Comprehensive Chemometrics**. [s.l.] Elsevier, 2020. v. 8p. 249–293.

GAO, T. *et al.* SPXYE: an improved method for partitioning training and validation sets. **Cluster Computing**, v. 22, n. s2, p. 3069–3078, 2019.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, p. 1–17, 1986.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2006.

GÓMEZ-CEBRIÁN, N. *et al.* Metabolomics contributions to the discovery of prostate cancer biomarkers. **Metabolites**, v. 9, n. 3, 2019.

GOUVEIA, L. R. **Metabonômica Aplicada ao Diagnóstico Diferencial de Doenças Hepáticas**. [s.l.] Universidade Federal de Pernambuco, 2017.

GRASMANN, G. *et al.* Gluconeogenesis in cancer cells – Repurposing of a starvation-induced metabolic pathway? **Biochimica et Biophysica Acta - Reviews on Cancer**, v. 1872, n. 1, p. 24–36, 2019.

GRIFFIN, J. L. Twenty years of metabonomics: so what has metabonomics done for toxicology? **Xenobiotica**, v. 50, n. 1, p. 110–114, 2020.

GUHA, R. *et al.* Deluge based Genetic Algorithm for feature selection. **Evolutionary Intelligence**, v. 14, n. 2, p. 357–367, 2021.

GÜNDOĞDU, S. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. **Multimedia Tools and Applications**, v. 82, n. 22, p. 34163–34181, 2023.

GUSTAFSSON, O.; MANSOUR, E. Prostate-specific Antigen (PSA), PSA Density and Age-adjusted PSA Reference Values in Screening for Prostate Cancer: A Study of a Randomly Selected Population of 2,400 Men. **Scandinavian Journal of Urology and Nephrology**, v. 32, n. 6, p. 373–377, 9 jan. 1998.

HACHCHAM, A. **XGBoost: Everything You Need to Know**.

HARRIS, C. R. *et al.* Array programming with NumPy. **Nature**, v. 585, n. 7825, p. 357–362, 2020.

HIRA, Z. M.; GILLIES, D. F. A review of feature selection and feature extraction methods applied on microarray data. **Advances in Bioinformatics**, v. 2015, n. 1, 2015.

HOGAN, C. A. *et al.* Nasopharyngeal metabolomics and machine learning approach for the diagnosis of influenza. **EBioMedicine**, v. 71, p. 103546, 2021.

HOTELLING, H. Analysis of a complex of statistical variables into Principal Components. Jour. Educ. Psych., 24, 417-441, 498-520. **The Journal of Educational Psychology**, v. 24, p. 417–441, 1933.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science \& Engineering**, v. 9, n. 3, p. 90–95, 2007.

INCA. **Estimativa 2020**. Disponível em: <https://www.inca.gov.br/estimativa/sintese-de-resultados-e-comentarios>. Acesso em: 20 mar. 2020.

JAURILA, H. *et al.* 1H NMR based metabolomics in human sepsis and healthy serum. **Metabolites**, v. 10, n. 2, p. 1–13, 2020.

JIA, W. *et al.* Feature dimensionality reduction: a review. **Complex and Intelligent Systems**, v. 8, n. 3, p. 2663–2693, 2022.

KAMITANI, R. *et al.* Evaluation of Gleason Grade Group 5 in a Contemporary Prostate Cancer Grading System and Literature Review. **Clinical Genitourinary Cancer**, v. 19, n. 1, p. 69-75.e5, fev. 2021.

KARAYANNIS, M. I.; EFSTATHIOU, C. E. Significant steps in the evolution of analytical chemistry - Is the todayCloseCurlys analytical chemistry only chemistry? **Talanta**, v. 102, p. 7–15, 2012.

KELLY, R. S. *et al.* Metabolomic Biomarkers of Prostate Cancer: Prediction, Diagnosis, Progression, Prognosis, and Recurrence. **Cancer Epidemiology Biomarkers & Prevention**, v. 25, n. 6, p. 887–906, 1 jun. 2016.

KHAIRE, U. M.; DHANALAKSHMI, R. Stability of feature selection algorithm: A review. **Journal of King Saud University - Computer and Information Sciences**, v. 34, n. 4, p. 1060–1073, 2022.

KO, S.; CHOI, J.; AHN, J. GVES: machine learning model for identification of prognostic genes with a small dataset. **Scientific Reports**, v. 11, n. 1, p. 1–8, 2021.

KORNYO, O. *et al.* Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms. **Computers and Security**, v. 135, n. August, p. 103456, 2023.

KOWALSKI, B.; BROWN, S.; VANDEGINSTE, B. Editorial. **Journal of Chemometrics**, v. 1, n. 1, p. 1–2, 1987.

KOWALSKI, B. R. Chemometrics: Views and Propositions. **Journal of Chemical Information and Computer Sciences**, v. 15, n. 4, p. 201–203, 1 nov. 1975.

KRZYWINSKI, M.; ALTMAN, N. Classification and regression trees. **Nature Methods**, v. 14, n. 8, p. 757–758, 2017.

KUPČE, Ē. *et al.* Parallel nuclear magnetic resonance spectroscopy. **Nature Reviews Methods Primers**, v. 1, n. 1, p. 27, 8 abr. 2021.

KWIATKOWSKI, R. **Gradient Descent Algorithm — a deep dive**.

LAVINE, B. K. Special Issue: Chemometrics. **Applied Spectroscopy**, v. 72, n. 3, p. 339, 2018.

LAVINE, B. K.; WORKMAN, J. Chemometrics. **Analytical Chemistry**, v. 85, n. 2, p. 705–714, 2013.

LAY, D. C. Matrizes Simétricas e Formas Qudráticas. *Em*: **Álgebra Linear e Suas Aplicações**. 2. ed. Rio de Janeiro, RJ: LTC, 2011. p. 504.

LEDOIT, O.; WOLF, M. Honey, I shrunk the sample covariance matrix. **Journal of Portfolio Management**, v. 30, n. 4, p. 1–22, 2004.

LEE, Y. R. *et al.* Untargeted Metabolomics and Steroid Signatures in Urine of Male Pattern Baldness Patients after Finasteride Treatment for a Year. **Metabolites**, v. 10, n. 4, p. 131, 30 mar. 2020.

LEMAITRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. 21 set. 2016.

LERCHE, M. H. *et al.* **NMR-Based Metabolomics**. New York, NY: Springer New York, 2019. v. 2037

LIN, L. S. *et al.* An attribute extending method to improve learning performance for small datasets. **Neurocomputing**, v. 286, p. 75–87, 2018.

LIN, L. S.; LIN, Y. S.; LI, D. C. Generating virtual samples to improve learning performance in small datasets with non-linear and asymmetric distributions. **Neurocomputing**, v. 548, p. 126408, 2023.

LIU, M.; MAO, X. Solvent Supression Methods in NMR Spectroscopy. *Em*: **Encyclopedia of Spectroscopy and Spectrometry**. [s.l.] Elsevier, 1999. v. 3p. 2604–2610.

LOH, W. Classification and regression trees. **WIREs Data Mining and Knowledge Discovery**, v. 1, n. 1, p. 14–23, 6 jan. 2011.

LOU, C.; ATOUI, M. A.; LI, X. **A new kernel trick embedded discriminant model for fault detection and diagnosis**2021 33rd Chinese Control and Decision Conference (CCDC). **Anais**...IEEE, 22 maio 2021

LUCAS, L. H. *et al.* Progress toward automated metabolic profiling of human serum: Comparison of CPMG and gradient-filtered NMR analytical methods. **Journal of Pharmaceutical and Biomedical Analysis**, v. 39, n. 1–2, p. 156–163, set. 2005.

LÜPSEN, H. Generalizations of the Tests by Kruskal-Wallis, Friedman and van der Waerden for Split-plot Designs. **Austrian Journal of Statistics**, v. 52, n. 5, p. 101–130, 11 set. 2023.

MAHESWARA RAO, V. V. R. *et al.* An Innovative Machine Learning based Heart Disease Assessment System by Sequential Feature Selection Approach. **2023 3rd International Conference on Intelligent Technologies, CONIT 2023**, n. Ml, p. 1–7, 2023.

MARAY, N. *et al.* Transfer Learning on Small Datasets for Improved Fall Detection. **Sensors**, v. 23, n. 3, 2023.

MCCLENATHAN, B. M. *et al.* Metabolites as biomarkers of adverse reactions following vaccination: A pilot study using nuclear magnetic resonance metabolomics. **Vaccine**, v. 35, n. 9, p. 1238–1245, mar. 2017.

MCKINNEY, W. **Data Structures for Statistical Computing in Python**2010

MENESES, P. R. *et al.* **Introdução ao Processamento de Imagens de Sensoriamento Remoto**. Brasília: UNB, 2012.

NAGANA GOWDA, G. A.; GOWDA, Y. N.; RAFTERY, D. Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. **Analytical Chemistry**, v. 87, n. 1, p. 706–715, 2015.

NANGA, S. *et al.* Review of Dimension Reduction Methods. **Journal of Data Analysis and Information Processing**, v. 09, n. 03, p. 189–231, 2021.

NETO, F. T. L. *et al.* 1H NMR-based metabonomics for infertility diagnosis in men with varicocele. **Journal of Assisted Reproduction and Genetics**, v. 37, n. 9, p. 2233–2247, 2020.

___. Prediction of semen analysis parameter improvement after varicocoelectomy using 1 H NMR-based metabonomics assays. **Andrology**, v. 10, n. 8, p. 1581–1592, 1 nov. 2022.

NICHOLSON, J. K. *et al.* 750 MHz 1H and 1H-13C NMR Spectroscopy of Human Blood Plasma. **Analytical Chemistry**, v. 67, n. 5, p. 793–811, 1995.

NICHOLSON, J. K.; BUCKINGHAM, M. J.; SADLER, P. J. High resolution 1H n.m.r. studies of vertebrate blood and plasma. **Biochemical Journal**, v. 211, n. 3, p. 605–615, 1983.

NICHOLSON, J. K.; LINDON, J. C. Metabonomics. **Nature**, v. 455, n. 7216, p. 1054–1056, 22 out. 2008.

NICHOLSON, J. K.; LINDON, J. C.; HOLMES, E. "Metabonomics": Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. **Xenobiotica**, v. 29, n. 11, p. 1181–1189, 1999.

OLIVEIRA, M. F. *et al.* Performance evaluate of different chemometrics formalisms used for prostate cancer diagnosis by NMR-based metabolomics. **Metabolomics**, v. 20, n. 1, p. 8, 21 dez. 2023.

PANDEY, S. K.; JANGHEL, R. R. Automatic detection of arrhythmia from imbalanced ECG database using CNN model with SMOTE. **Australasian Physical and Engineering Sciences in Medicine**, v. 42, n. 4, p. 1129–1139, 2019.

PEDREGOSA, F. *et al.* **Scikit-Learn**. Disponível em: <https://scikit-learn.org/stable>. Acesso em: 9 jan. 2024.

PEDREGOSA, FABIAN *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011.

PÉREZ-RAMBLA, C. *et al.* Non-invasive urinary metabolomic profiling discriminates prostate cancer from benign prostatic hyperplasia. **Metabolomics**, v. 13, n. 5, p. 1–12, 2017.

PINTO, R. C. Chemometrics Methods and Strategies in Metabolomics. *Em*: [s.l: s.n.]. v. 965p. 163–190.

___. Chemometrics Methods and Strategies in Metabolomics. *Em*: [s.l: s.n.]. v. 965p. 163–190.

POPOVIC, A. *et al.* Review of the most common chemometric techniques in illicit drug profiling. **Forensic Science International**, v. 302, p. 109911, 2019.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning - Third Edition**. 3rd. ed. Birmigham, UK: Packt Publishing, 2019.

RAWLA, P. Epidemiology of Prostate Cancer. **World Journal of Oncology**, v. 10, n. 2, p. 63–89, 2019.

REENEN, M. VAN *et al.* Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp. **BMC Bioinformatics**, v. 18, n. 1, p. 1–13, 2017.

REMMERS, S. *et al.* Improving the prediction of biochemical recurrence after radical prostatectomy with the addition of detailed pathology of the positive surgical margin and cribriform growth. **Annals of Diagnostic Pathology**, v. 56, p. 151842, fev. 2022.

RODIONOVA, O. Y.; TITOVA, A. V.; POMERANTSEV, A. L. Discriminant analysis is an inappropriate method of authentication. **TrAC - Trends in Analytical Chemistry**, v. 78, p. 17–22, 2016.

RODRIGUES, A. DE P.; LUNA, A. S.; PINTO, L. An evaluation strategy to select and discard sampling preprocessing methods for imbalanced datasets: A focus on classification models. **Chemometrics and Intelligent Laboratory Systems**, v. 240, n. August, p. 104933, 2023.

RODRIGUEZ-TORRES, F.; CARRASCO-OCHOA, J. A.; MARTÍNEZ-TRINIDAD, J. F. Deterministic oversampling methods based on SMOTE. **Journal of Intelligent and Fuzzy Systems**, v. 36, n. 5, p. 4945–4955, 2019.

ROSS, A. *et al.* NMR Spectroscopy Techniques for Application to Metabonomics. *Em*: **The Handbook of Metabonomics and Metabolomics**. [s.l.] Elsevier, 2007. p. 55–112.

SADEGHIAN, Z. *et al.* A review of feature selection methods based on meta-heuristic algorithms. **Journal of Experimental and Theoretical Artificial Intelligence**, v. 00, n. 00, p. 1–51, 2023.

SHEN, J. *et al.* A Hybrid Method to Predict Postoperative Survival of Lung Cancer Using Improved SMOTE and Adaptive SVM. **Computational and Mathematical Methods in Medicine**, v. 2021, 2021.

SILVA, R. D. DA. **Aplicações Metabonômicas usando Ressonância Magnética Nuclear de 1H: Diagnóstico Não-Invasivo de Câncer de Próstata e Urológico & Classificação de Azeite de Oliva Extra Virgem de Produção Orgânica**. [s.l.] Universidade Federal de Pernambuco, 2017.

SIMON, N. I. *et al.* Best Approaches and Updates for Prostate Cancer Biochemical Recurrence. **American Society of Clinical Oncology Educational Book**, n. 42, p. 352–359, jul. 2022.

SOLOMONS, T. W. G.; FRYHLE, C. B. Ressonância Magnética Nuclear e Espectrometria de Massas: Ferramentas para a Determinação de Estrutura. *Em*: **Química Orgânica, 1**. 8ª ed. Rio de Janeiro, RJ: LTC, 2005. .

SPEISER, J. L. *et al.* A comparison of random forest variable selection methods for classification prediction modeling. **Expert Systems with Applications**, v. 134, p. 93–101, 2019.

SREEJITH, S.; KHANNA NEHEMIAH, H.; KANNAN, A. Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. **Computers in Biology and Medicine**, v. 126, 1 nov. 2020.

STABILE, A. *et al.* Multiparametric MRI for prostate cancer diagnosis: current status and future directions. **Nature Reviews Urology**, v. 17, n. 1, p. 41–61, 2020.

STAGLJAR, I. The power of OMICs. **Biochemical and Biophysical Research Communications**, v. 479, n. 4, p. 607–609, 2016.

SUD, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. **Nucleic Acids Research**, v. 44, n. D1, p. D463–D470, 4 jan. 2016.

SUNG, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. **CA: A Cancer Journal for Clinicians**, v. 71, n. 3, p. 209–249, 4 maio 2021.

TAYLOR, P.; KENNARD, R. W.; STONE, L. A. Technometrics Computer Aided Design of Experiments. **Technometric**, v. 11, n. 1, p. 137–148, 1969.

THARWAT, A. *et al.* Linear discriminant analysis: A detailed tutorial. **AI Communications**, v. 30, n. 2, p. 169–190, 2017.

TOTH, R. *et al.* Random forest-based modelling to detect biomarkers for prostate cancer progression. **bioRxiv**, p. 1–15, 2019.

TOURINHO-BARBOSA, R. R.; POMPEO, A. C. L.; GLINA, S. Prostate cancer in Brazil and Latin America: Epidemiology and screening. **International Braz J Urol**, v. 42, n. 6, p. 1081–1090, 2016.

TRYGG, J.; WOLD, S. Orthogonal projections to latent structures (O-PLS). **Journal of Chemometrics**, v. 16, n. 3, p. 119–128, 2002.

UMER, M. *et al.* Scientific papers citation analysis using textual features and SMOTE resampling techniques. **Pattern Recognition Letters**, v. 150, p. 250–257, 2021.

VABALAS, A. *et al.* Machine learning algorithm validation with a limited sample size. **PLoS ONE**, v. 14, n. 11, p. 1–20, 2019.

VACH, W.; GERKE, O. Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties. **MethodsX**, v. 10, p. 102212, 2023.

VANDERGRIFT, L. A. *et al.* Metabolomic Prediction of Human Prostate Cancer Aggressiveness: Magnetic Resonance Spectroscopy of Histologically Benign Tissue. **Scientific Reports**, v. 8, n. 1, p. 1–12, 2018.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. New York, NY: Springer New York, 2000.

VIRTANEN, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, n. 3, p. 261–272, 2020.

VISHRAJ, R.; GUPTA, S.; SINGH, S. Evaluation of feature selection methods utilizing random forest and logistic regression for lung tissue categorization using HRCT images. **Expert Systems**, v. 40, n. 8, p. 1–30, 2023.

WANG, S. *et al.* Research on expansion and classification of imbalanced data based on SMOTE algorithm. **Scientific Reports**, v. 11, n. 1, p. 1–11, 2021.

WANG, Z.; DONG, C. Gluconeogenesis in Cancer: Function and Regulation of PEPCK, FBPase, and G6Pase. **Trends in Cancer**, v. 5, n. 1, p. 30–45, 2019.

WASKOM, M. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 6 abr. 2021.

WEI, Y. *et al.* Multi-scale sequential feature selection for disease classification using Raman spectroscopy data. **Computers in Biology and Medicine**, v. 162, n. April, 2023.

WILIMITIS, D. **The Kernel Trick in Support Vector Classification**.

WU, Y.; FANG, Y. Stroke prediction with machine learning methods among older chinese. **International Journal of Environmental Research and Public Health**, v. 17, n. 6, 2 mar. 2020.

XU, B. *et al.* Metabolomics Profiling Discriminates Prostate Cancer From Benign Prostatic Hyperplasia Within the Prostate-Specific Antigen Gray Zone. **Frontiers in Oncology**, v. 11, n. October, p. 1–13, 2021.

YAN, X.; ZHU, H. A novel robust support vector machine classifier with feature mapping. **Knowledge-Based Systems**, v. 257, p. 109928, 2022.

ZHANG, Y. *et al.* Metabolomics approach by 1 H NMR spectroscopy of serum reveals progression axes for asymptomatic hyperuricemia and gout. **Arthritis Research and Therapy**, v. 20, n. 1, p. 1–11, 2018.

ZHANG, Z. *et al.* Deep learning in omics: A survey and guideline. **Briefings in Functional Genomics**, v. 18, n. 1, p. 41–57, 2019.

ZHAO, L.-L. *et al.* NMR Metabolomics and Random Forests Models to Identify Potential Plasma Biomarkers of Blood Stasis Syndrome With Coronary Heart Disease Patients. **Frontiers in Physiology**, v. 10, 4 set. 2019.

ZHAO, W. *et al.* XGB model: Research on evaporation duct height prediction based on XGBoost algorithm. **Radioengineering**, v. 29, n. 1, p. 81–93, 2020.

ZHENG, H. *et al.* NMR-based metabolomics analysis identifies discriminatory metabolic disturbances in tissue and biofluid samples for progressive prostate cancer. **Clinica Chimica Acta**, v. 501, n. October 2019, p. 241–251, 2020.

APÊNDICE A – RESULTADOS PARA O CONJUNTO DE DADOS DE DIAGNÓSTICO

Tabela A1. Resultados do laço principal do trabalho, incluindo o tempo em segundos e os parâmetros de cada estimador. Em negrito os melhores resultados.

Seletor	Est.	Ex.	Sens.	Esp.	AUROC	Kappa	Tempo	VS	Parâmetros
NA	SVM	0.64	1.00	0.25	0.63	0.26	6.42	NA	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
NA	LDA	0.72	0.85	0.58	0.71	0.43	9.64	NA	{'shrinkage': 'auto', 'solver': 'eigen'}
NA	LR	0.72	0.92	0.50	0.71	0.43	10.21	NA	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
NA	KNN	0.80	0.85	0.75	0.80	0.60	11.39	NA	{'n_neighbors': 6, 'p': 1, 'weights': 'distance'}
NA	XGB	0.52	1.00	0.00	0.50	0.00	424.06	NA	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFM-LR	SVM	0.64	1.00	0.25	0.63	0.26	0.25	50	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFM-LR	LDA	0.60	1.00	0.17	0.58	0.17	0.31	50	{'shrinkage': 'auto', 'solver': 'eigen'}
SFM-LR	LR	0.64	1.00	0.25	0.63	0.26	0.52	50	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFM-LR	KNN	0.84	1.00	0.67	0.83	0.68	0.86	50	{'n_neighbors': 5, 'p': 2, 'weights': 'distance'}
SFM-LR	XGB	0.52	1.00	0.00	0.50	0.00	91.58	50	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFM-ETC	SVM	0.72	0.92	0.50	0.71	0.43	0.20	9	{'C': 10, 'dual': False, 'penalty': 'l2', 'random_state': 123}
SFM-ETC	LDA	0.76	0.85	0.67	0.76	0.52	0.25	9	{'shrinkage': 'auto', 'solver': 'eigen'}
SFM-ETC	LR	0.72	0.92	0.50	0.71	0.43	0.43	9	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFM-ETC	KNN	0.88	1.00	0.75	0.88	0.76	0.77	9	{'n_neighbors': 6, 'p': 1, 'weights': 'distance'}
SFM-ETC	XGB	0.56	0.85	0.25	0.55	0.10	78.08	9	{'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0.5, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
RFECV-LR	SVM	0.56	1.00	0.08	0.54	0.09	81.77	4	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
RFECV-LR	LDA	0.60	1.00	0.17	0.58	0.17	81.81	4	{'shrinkage': 'auto', 'solver': 'eigen'}
RFECV-LR	LR	0.60	1.00	0.17	0.58	0.17	81.99	4	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
RFECV-LR	KNN	0.72	1.00	0.42	0.71	0.43	82.35	4	{'n_neighbors': 2, 'p': 1, 'weights': 'uniform'}

Tabela A1. (continuação)

Seletor	Est.	Ex.	Sens.	Esp.	AUROC	Kappa	Tempo	VS	Parâmetros
RFECV-LR	XGB	0.52	1.00	0.00	0.50	0.00	158.09	4	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
RFECV-ETC	SVM	0.72	0.85	0.58	0.71	0.43	7.80	507	{'C': 0.001, 'dual': False, 'penalty': 'l2', 'random_state': 123}
RFECV-ETC	LDA	0.76	0.77	0.75	0.76	0.52	8.41	507	{'shrinkage': 'auto', 'solver': 'eigen'}
RFECV-ETC	LR	0.72	0.92	0.50	0.71	0.43	8.76	507	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
RFECV-ETC	KNN	0.68	0.62	0.75	0.68	0.36	9.26	507	{'n_neighbors': 7, 'p': 1, 'weights': 'distance'}
RFECV-ETC	XGB	0.72	0.92	0.50	0.71	0.43	306.65	507	{'gamma': 0, 'learning_rate': 0.3, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 1, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFS-LR	SVM	0.56	1.00	0.08	0.54	0.09	861.55	24	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFS-LR	LDA	0.60	1.00	0.17	0.58	0.17	861.64	24	{'shrinkage': 'auto', 'solver': 'eigen'}
SFS-LR	LR	0.60	1.00	0.17	0.58	0.17	861.84	24	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFS-LR	KNN	0.80	1.00	0.58	0.79	0.59	862.29	24	{'n_neighbors': 6, 'p': 3, 'weights': 'distance'}
SFS-LR	XGB	0.52	1.00	0.00	0.50	0.00	942.13	24	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFS-ETC	SVM	0.68	1.00	0.33	0.67	0.34	191.49	24	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFS-ETC	LDA	0.68	1.00	0.33	0.67	0.34	191.55	24	{'shrinkage': 'auto', 'solver': 'eigen'}
SFS-ETC	LR	0.76	1.00	0.50	0.75	0.51	191.74	24	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFS-ETC	KNN	0.80	0.92	0.67	0.79	0.60	192.07	24	{'n_neighbors': 7, 'p': 3, 'weights': 'distance'}
SFS-ETC	XGB	0.52	1.00	0.00	0.50	0.00	272.19	24	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
GA	**SVM**	**0.80**	**0.85**	**0.75**	**0.80**	**0.60**	**77.14**	**6**	**{'C': 0.001, 'dual': False, 'penalty': 'l1', 'random_state': 123}**
GA	**LDA**	**0.88**	**0.92**	**0.83**	**0.88**	**0.76**	**116.41**	**5**	**{'shrinkage': 'auto', 'solver': 'eigen'}**
GA	**LR**	**0.88**	**1.00**	**0.75**	**0.88**	**0.76**	**179.36**	**5**	**{'C': 0.05, 'max_iter': 100, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**

APÊNDICE B – RESULTADOS PARA O CONJUNTO DE DADOS PARA DIAGNÓSTICO DE ASMA EM GATOS

Tabela B1. Resultados para o conjunto de dados para diagnóstico de asma em gatos, incluindo o tempo em segundos e os parâmetros de cada estimador. Em negrito os melhores resultados.

Seletor	Est.	Ex.	Sens.	Esp.	AUROC	Kappa	Tempo	VS	Parâmetros
NA	SVM	0.84	0.74	1.00	0.87	0.69	17.90	NA	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
NA	LDA	0.91	1.00	0.77	0.88	0.80	326.11	NA	{'shrinkage': 'auto', 'solver': 'eigen'}
NA	LR	0.69	1.00	0.23	0.62	0.26	328.16	NA	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
NA	KNN	0.72	0.95	0.38	0.67	0.36	331.45	NA	{'n_neighbors': 2, 'p': 3, 'weights': 'distance'}
NA	XGB	0.66	0.42	1.00	0.71	0.37	5055.01	NA	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFM-LR	SVM	0.84	0.74	1.00	0.87	0.69	0.47	50	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFM-LR	LDA	0.88	0.79	1.00	0.89	0.75	0.54	50	{'shrinkage': 'auto', 'solver': 'eigen'}
SFM-LR	LR	0.84	0.74	1.00	0.87	0.69	0.75	50	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFM-LR	KNN	0.84	0.74	1.00	0.87	0.69	1.12	50	{'n_neighbors': 6, 'p': 3, 'weights': 'distance'}
SFM-LR	XGB	0.66	0.42	1.00	0.71	0.37	103.46	50	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFM-ETC	SVM	0.41	0.21	0.69	0.45	-0.09	0.18	16	{'C': 0.1, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFM-ETC	LDA	0.41	0.16	0.77	0.46	-0.06	0.23	16	{'shrinkage': None, 'solver': 'eigen'}
SFM-ETC	LR	0.38	0.11	0.77	0.44	-0.11	0.42	16	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFM-ETC	KNN	0.72	0.89	0.46	0.68	0.38	0.77	16	{'n_neighbors': 2, 'p': 1, 'weights': 'distance'}
SFM-ETC	XGB	0.50	0.53	0.46	0.49	-0.01	90.43	16	{'gamma': 5, 'learning_rate': 0.1, 'max_depth': 10, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0.5, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
RFECV-LR	SVM	0.91	0.84	1.00	0.92	0.81	1997.88	617	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
RFECV-LR	**LDA**	**0.97**	**0.95**	**1.00**	**0.97**	**0.94**	**1998.82**	**617**	**{'shrinkage': 'auto', 'solver': 'eigen'}**
RFECV-LR	**LR**	**0.97**	**1.00**	**0.92**	**0.96**	**0.93**	**1999.19**	**617**	**{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**
RFECV-LR	KNN	0.84	0.95	0.69	0.82	0.66	1999.97	617	{'n_neighbors': 6, 'p': 2, 'weights': 'distance'}

Tabela B1. (continuação)

Seletor	Est.	Ex.	Sens.	Esp.	AUROC	Kappa	Tempo	VS	Parâmetros
RFECV-LR	XGB	0.66	0.42	1.00	0.71	0.37	2438.71	617	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
RFECV-ETC	SVM	0.94	1.00	0.85	0.92	0.87	87.57	1565	{'C': 1, 'dual': False, 'penalty': 'l1', 'random_state': 123}
RFECV-ETC	LDA	0.94	1.00	0.85	0.92	0.87	97.64	1565	{'shrinkage': 'auto', 'solver': 'eigen'}
RFECV-ETC	LR	0.78	1.00	0.46	0.73	0.50	98.35	1565	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
RFECV-ETC	KNN	0.59	0.84	0.23	0.54	0.08	99.67	1565	{'n_neighbors': 2, 'p': 2, 'weights': 'distance'}
RFECV-ETC	**XGB**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1246.27**	**1565**	**{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}**
SFS-LR	SVM	0.88	0.79	1.00	0.89	0.75	9115.36	24	{'C': 1, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFS-LR	LDA	0.81	0.68	1.00	0.84	0.64	9115.43	24	{'shrinkage': 'auto', 'solver': 'eigen'}
SFS-LR	LR	0.88	0.79	1.00	0.89	0.75	9115.61	24	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFS-LR	KNN	0.84	0.74	1.00	0.87	0.69	9116.06	24	{'n_neighbors': 3, 'p': 1, 'weights': 'distance'}
SFS-LR	XGB	0.66	0.42	1.00	0.71	0.37	9205.20	24	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFS-ETC	SVM	0.84	0.74	1.00	0.87	0.69	1198.17	24	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFS-ETC	LDA	0.84	0.74	1.00	0.87	0.69	1198.23	24	{'shrinkage': 'auto', 'solver': 'eigen'}
SFS-ETC	LR	0.84	0.74	1.00	0.87	0.69	1198.41	24	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFS-ETC	KNN	0.81	0.68	1.00	0.84	0.64	1198.84	24	{'n_neighbors': 4, 'p': 1, 'weights': 'distance'}
SFS-ETC	XGB	0.66	0.42	1.00	0.71	0.37	1287.65	24	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
GA	**SVM**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**305.85**	**1**	**{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}**
GA	LDA	0.81	0.68	1.00	0.84	0.64	628.88	3	{'shrinkage': 'auto', 'solver': 'eigen'}
GA	**LR**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**916.58**	**2**	**{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**

APÊNDICE C – RESULTADOS PARA O CONJUNTO DE DADOS ACERCA DA IMUNIZAÇÃO DA VARÍOLA

Tabela C1. Resultados para o conjunto de dados acerca da imunização da varíola, incluindo o tempo em segundos e os parâmetros de cada estimador. Em negrito os melhores resultados.

Seletor	Est.	Ex.	Sens.	Esp.	AUROC	Kappa	Tempo	VS	Parâmetros
NA	SVM	0.77	0.65	1.00	0.83	0.56	4.18	NA	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
NA	LDA	0.74	0.61	1.00	0.80	0.52	4.36	NA	{'shrinkage': 'auto', 'solver': 'eigen'}
NA	**LR**	**0.80**	**0.70**	**1.00**	**0.85**	**0.61**	**4.78**	**NA**	**{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**
NA	KNN	0.57	0.57	0.58	0.57	0.14	5.68	NA	{'n_neighbors': 3, 'p': 2, 'weights': 'distance'}
NA	XGB	0.74	0.61	1.00	0.80	0.52	187.79	NA	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFM-LR	SVM	0.74	0.61	1.00	0.80	0.52	0.23	46	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFM-LR	LDA	0.69	0.52	1.00	0.76	0.43	0.29	46	{'shrinkage': 'auto', 'solver': 'eigen'}
SFM-LR	LR	0.74	0.61	1.00	0.80	0.52	0.50	46	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFM-LR	KNN	0.66	0.48	1.00	0.74	0.39	0.84	46	{'n_neighbors': 5, 'p': 3, 'weights': 'distance'}
SFM-LR	XGB	0.74	0.61	1.00	0.80	0.52	102.53	46	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFM-ETC	SVM	0.74	0.61	1.00	0.80	0.52	0.17	5	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
SFM-ETC	LDA	0.60	0.39	1.00	0.70	0.31	0.23	5	{'shrinkage': 'auto', 'solver': 'eigen'}
SFM-ETC	LR	0.71	0.57	1.00	0.78	0.47	0.39	5	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFM-ETC	KNN	0.60	0.39	1.00	0.70	0.31	0.71	5	{'n_neighbors': 4, 'p': 3, 'weights': 'distance'}
SFM-ETC	XGB	0.74	0.61	1.00	0.80	0.52	74.89	5	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
RFECV-LR	**SVM**	**0.80**	**0.70**	**1.00**	**0.85**	**0.61**	**21.43**	**8**	**{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}**
RFECV-LR	LDA	0.63	0.43	1.00	0.72	0.35	21.49	8	{'shrinkage': 'auto', 'solver': 'eigen'}
RFECV-LR	LR	0.74	0.61	1.00	0.80	0.52	21.65	8	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
RFECV-LR	KNN	0.57	0.35	1.00	0.67	0.27	21.97	8	{'n_neighbors': 6, 'p': 1, 'weights': 'distance'}

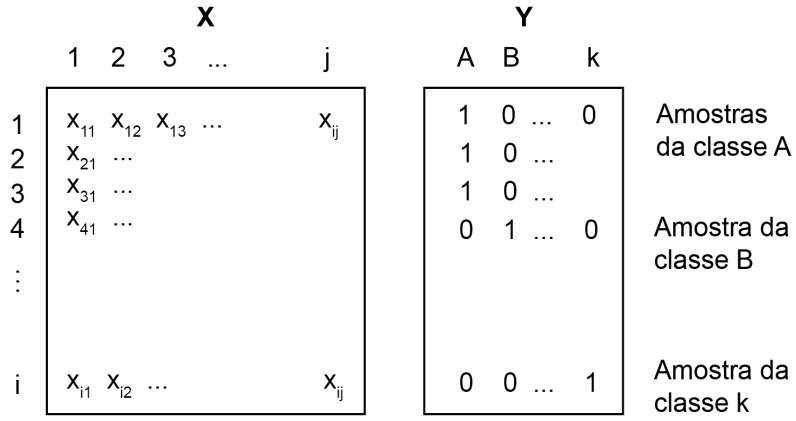
Tabela C1. (continuação)

Seletor	Est.	Ex.	Sens.	Esp.	AUROC	Kappa	Tempo	VS	Parâmetros
RFECV-LR	XGB	0.74	0.61	1.00	0.80	0.52	137.31	8	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
RFECV-ETC	SVM	0.77	0.65	1.00	0.83	0.56	2.00	185	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
RFECV-ETC	LDA	0.74	0.61	1.00	0.80	0.52	2.13	185	{'shrinkage': 'auto', 'solver': 'eigen'}
RFECV-ETC	**LR**	**0.80**	**0.70**	**1.00**	**0.85**	**0.61**	**2.42**	**185**	**{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**
RFECV-ETC	KNN	0.57	0.57	0.58	0.57	0.14	2.94	185	{'n_neighbors': 3, 'p': 2, 'weights': 'distance'}
RFECV-ETC	XGB	0.74	0.61	1.00	0.80	0.52	275.04	185	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFS-LR	SVM	0.74	0.61	1.00	0.80	0.52	294.27	24	{'C': 10, 'dual': False, 'penalty': 'l2', 'random_state': 123}
SFS-LR	LDA	0.63	0.43	1.00	0.72	0.35	294.32	24	{'shrinkage': 'auto', 'solver': 'eigen'}
SFS-LR	LR	0.74	0.61	1.00	0.80	0.52	294.53	24	{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}
SFS-LR	KNN	0.69	0.52	1.00	0.76	0.43	294.88	24	{'n_neighbors': 4, 'p': 3, 'weights': 'distance'}
SFS-LR	XGB	0.74	0.61	1.00	0.80	0.52	386.15	24	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
SFS-ETC	SVM	0.80	0.70	1.00	0.85	0.61	40.46	24	{'C': 10, 'dual': False, 'penalty': 'l2', 'random_state': 123}
SFS-ETC	LDA	0.71	0.57	1.00	0.78	0.47	40.51	24	{'shrinkage': 'auto', 'solver': 'eigen'}
SFS-ETC	**LR**	**0.83**	**0.74**	**1.00**	**0.87**	**0.66**	**40.70**	**24**	**{'C': 10, 'max_iter': 500, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**
SFS-ETC	KNN	0.71	0.61	0.92	0.76	0.45	41.03	24	{'n_neighbors': 6, 'p': 2, 'weights': 'distance'}
SFS-ETC	XGB	0.74	0.61	1.00	0.80	0.52	128.64	24	{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 0, 'reg_alpha': 0, 'reg_lambda': 0, 'seed': 123, 'tree_method': 'hist', 'verbosity': 0}
GA	SVM	0.77	0.65	1.00	0.83	0.56	61.52	5	{'C': 10, 'dual': False, 'penalty': 'l1', 'random_state': 123}
GA	LDA	0.60	0.39	1.00	0.70	0.31	145.89	7	{'shrinkage': 'auto', 'solver': 'eigen'}
GA	**LR**	**0.80**	**0.70**	**1.00**	**0.85**	**0.61**	**281.44**	**7**	**{'C': 0.05, 'max_iter': 100, 'penalty': 'l2', 'random_state': 123, 'solver': 'lbfgs'}**

APÊNDICE D – ANÁLISE DISCRIMINANTE POR QUADRADOS MÍNIMOS PARCIAIS E PROJEÇÕES ORTOGONAIS PARA ESTRUTURAS LATENTES

O PLS-DA (do inglês, *Partial Least Squares Discriminant Analysis*) é bastante utilizado na análise multivariada e na ciência de dados em geral, como método de classificação. Assim como a PCA é um método de redução de dimensionalidade, onde os dados são ajustados linearmente das variáveis originais para um espaço de menor espaço dimensional. O método é derivado da regressão PLS, porém nesse caso, além da matriz de dados **X** (I x J), a matriz resposta **Y** (I x K) contém a informação das variáveis categóricas, ou seja, possui K colunas, para cada classe, que apresentam valores binários, 0 e 1, como indica a Figura D1 (Rodionova, Titova e Pomerantsev, 2016).

Figura D1. Matrizes para construção de modelo PLS-DA.



Fonte: Adaptado de Gouveia (2017).
Nota: Matrizes **X** e **Y** com i amostras, j variáveis e k classes.

Caso o número de classes, k = 2, pode se escrever **Y** como um vetor **y**, onde é assumido valor -1 para as amostras de uma classe e 1 para as amostras da outra classe.

O PLS-DA consiste em encontrarmos as matrizes de pesos e escores em um novo espaço amostral, cujo as novas variáveis são chamadas de Variáveis Latentes (do inglês: LV). As matrizes de pesos **L** e de escores **T** podem ser calculadas pelo método NIPALS, que nasceu juntamente com o surgimento da regressão PLS em 1977 (Geladi e Kowalski, 1986; Trygg e Wold, 2002). O algoritmo NIPALS (Equações D1-D12), proposto por Wold, para a construção de um modelo PLS está descrito no quadro abaixo (Quadro D1).

Quadro D1. Algoritmo NIPALS para a regressão PLS.

Pré-processamento das matrizes **X** e **y**			
Cálculo da primeira LV			
	$\mathbf{X} = \mathbf{y}\mathbf{w}_1^{\mathrm{T}} + \mathbf{E}$		(D1)
	$\widehat{\mathbf{w}}_1 = \mathbf{X}^{\mathrm{T}}\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1}$	(fator-peso)	(D2)
Matriz **X**	$\widehat{\mathbf{w}}_1 = \widehat{\mathbf{w}}_1 / \sqrt{\widehat{\mathbf{w}}_1^{\mathrm{T}}\widehat{\mathbf{w}}_1}$	(normalização)	(D3)
	$\mathbf{t}_1 = \mathbf{X}\widehat{\mathbf{w}}$	(escores)	(D4)
	$\hat{\mathbf{l}}_1 = \mathbf{X}^{\mathrm{T}}\mathbf{t}_1(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1)^{-1}$	(pesos)	(D5)
Vetor **y**	$\mathbf{y} = \mathrm{q}_1\boldsymbol{t}_1 + \mathbf{f}$	(regressão do vetor **y** no vetor $\mathbf{t}_1$)	(D6)
	$\mathrm{q}_1 = (\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1)^{-1}\boldsymbol{t}_1^{T}\boldsymbol{y}$		(D7)
Atualização de **X** e **Y**	$\mathbf{E} = \mathbf{X} - \mathbf{t}_1\mathbf{l}_1^{\mathrm{T}}$	(matrizes de erros)	(D8)
	$\mathbf{f} = \mathbf{y} - \mathbf{t}_1\mathrm{q}_1$		(D9)
	$\mathbf{X} = \mathbf{E}$		(D10)
	$\mathbf{y} = \mathbf{f}$		(D11)
Repetir os cálculos das demais LVs			
Cálculo do vetor de regressão	$\hat{\mathbf{b}} = \boldsymbol{W}(\boldsymbol{L}^{T}\boldsymbol{W})^{-1}\boldsymbol{q}$		(D12)

Fonte: Adaptado de Ferreira (2015, p. 349).

No algoritmo NIPALS, a primeira etapa (Equação D2) consiste no cálculo do vetor de fatores-peso $\hat{\mathbf{w}}_1$ para a primeira LV, por isso o índice 1, a partir de uma regressão linear entre as matrizes **X** e o vetor **y**. Esse vetor de regressão é normalizado (Equação D3) e então são calculados os escores (Equação D4), $\mathbf{t}_1$, que são as coordenadas de cada amostra no vetor $\hat{\mathbf{w}}_1$. Esses escores são utilizados para fazer a regressão linear dos pesos (Equação D5), $\hat{\mathbf{l}}_1$, e o coeficiente q_1, que é o coeficiente da regressão de **y** nos escores $\mathbf{t}_1$ (Equação D6). Então os erros **E** e **f** relativos às regressões da matriz de dados **X** e de resposta **y**, são calculados e são utilizados como as matrizes iniciais para as próximas LVs (Equações D7 – D11). Por fim é calculado o vetor de regressão $\hat{\mathbf{b}}$, utilizando a matriz **W** dos fatores-peso, a matriz de pesos **L** das variáveis latentes e o **q** o vetor dos coeficientes de regressão **y** em **T** (Equação D12) (Ferreira, 2015; Geladi e Kowalski, 1986).

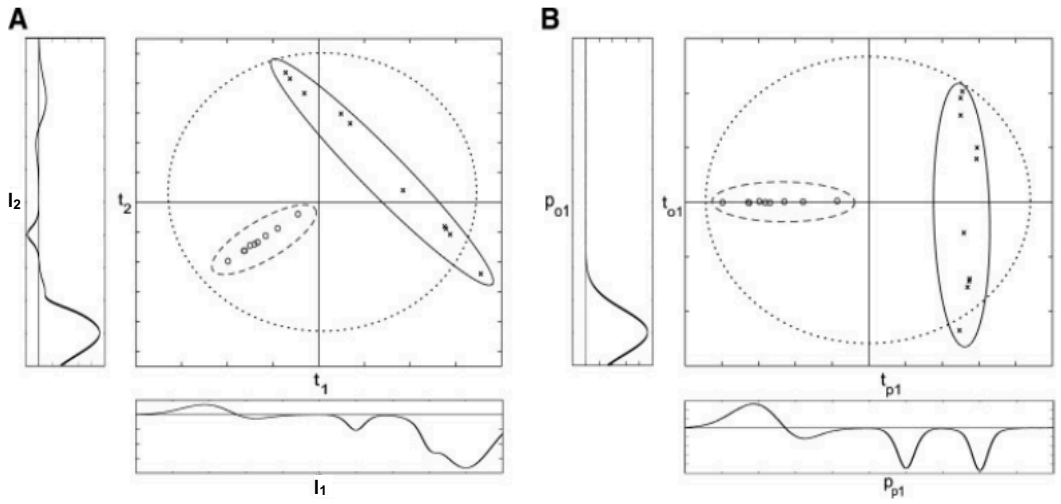
Sabendo-se que o método de PLS maximiza a relação entre a variável dependente e os escores (Equação D6), as variáveis latentes representam a direção no sentido da melhor discriminação entre as classes. Os valores do vetor resposta após a regressão não é exatamente os valores inteiros utilizados na matriz de entrada **Y**, porém deve apresentar valores próximos a esses.

Como citado anteriormente, a PLS-DA é um método largamente utilizado, principalmente pela sua adaptabilidade em situações de auto colinearidade entre as variáveis e

de ruídos. Porém, utilizando um filtro de correção ortogonal de sinal (do inglês: OSC), desenvolvido por Wold e seus colaboradores, pôde se chegar a uma modificação do PLS-DA chamada de OPLS-DA (Pinto, 2017b).

Na OPLS-DA, como já citado acima, utilizamos um filtro OSC, que tem a função de remover a variação sistemática da matriz **X** não correlacionada com a matriz resposta **Y**, sendo assim, o filtro remove a variabilidade em **X** que é ortogonal a **Y** (Trygg e Wold, 2002). Nesse caso, como ilustrado na Figura 14, pode-se observar que uma força do OPLS-DA é o seu poder de separar a variação preditiva entre as classes (Figura D2B), t_{p1}, da variação não preditiva, t_{o1}, que nesse caso está associada a variação entre as amostras da mesma classe. Para a PLS-DA, dois componentes t_1 e t_2 são necessários para encontrar um plano perfeitamente discriminatório entre as duas classes, conforme mostrado na Figura D2A. Os vetores de peso correspondentes l_1 e l_2 conterão uma mistura de ambas as propriedades discriminatórias, bem como as propriedades não discriminatórias que são principalmente confundidas com a direção de t_2 (Bylesjö *et al.*, 2006), diferentemente dos pesos p_{p1} e p_{o1} do OPLS-DA

Figura D2. Diferenças entre os modelos (A) PLS-DA e (B) OPLS-DA, num exemplo contendo duas classes.



Fonte: (Bylesjö *et al.*, 2006)

A Equação 31 demonstra que a matriz de dados **X** é descomposta em três termos.

$$\mathbf{X} = \mathbf{T}_p\mathbf{L}_p^T + \mathbf{T}_o\mathbf{L}_o^T + \mathbf{E} \quad \text{(D13)}$$

Onde, $\mathbf{T}_p$ representa a matriz de escores e $\mathbf{L}_p$ a matriz de pesos que predizem as classes e são correlacionadas a matriz resposta **Y**, e $\mathbf{T}_o$ e $\mathbf{L}_o$ as matrizes de escores e de pesos da ortogonal de **Y** e **E** é a matriz de resíduos do modelo. O método desenvolvido por Trygg e Wold (2002) para o OPLS-DA utiliza um método modificado de NIPALS para PLS (Quadro D1), calculando os valores de escores, pesos e fatores-peso para a coordenada ortogonal.

Além disso, a importância das variáveis para o PLS-DA são mostradas a partir de dois formalismos que são os escores da projeção de importância das variáveis (do inglês: VIP), que foi um método proposto por Wold e colaboradores. Esses escores são utilizados para selecionar variáveis de interesse.

O risco de sobreajuste pode ser reduzido com a utilização de um método de validação, principalmente quando não há um grande volume de amostras e não possa ser possível dividir amostras para teste. Existem diversos tipo de métodos de validação, a utilizada nesse documento é a *leave-one-out cross validation* (LOOCV), que consiste no cálculo do erro do modelo utilizando n - 1 amostras (onde n é o número de amostras), e testando a amostra restante. Este processo é repetido para todos os n subconjuntos de tamanho n - 1 (Gouveia, 2017; Trygg e Wold, 2002).

A LOOCV pode resultar no cálculo de dois fatores o R^2 (Equação D14) e o Q^2, que são, respectivamente, a qualidade do ajuste e a habilidade preditiva do modelo. Baixos valores de Q^2 podem demonstrar que o modelo está sobreajustado ou possuam baixo poder de predição (Chong, Wishart e Xia, 2019). O Q^2 pode ser calculado a partir da Equação D17, e é definido como 1 menos o quociente entre a soma quadrática dos resíduos de predição (do inglês: PRESS) e a soma quadrática total (do inglês: TSS), que se apresentam nas Equações D15 e D16.

$$R^2 = \frac{SQ_{reg}}{SQ_T - SQ_M} \tag{D14}$$

$$PRESS = \sum_{n=1}^{I} (y_i - \hat{y}_i)^2 \tag{D15}$$

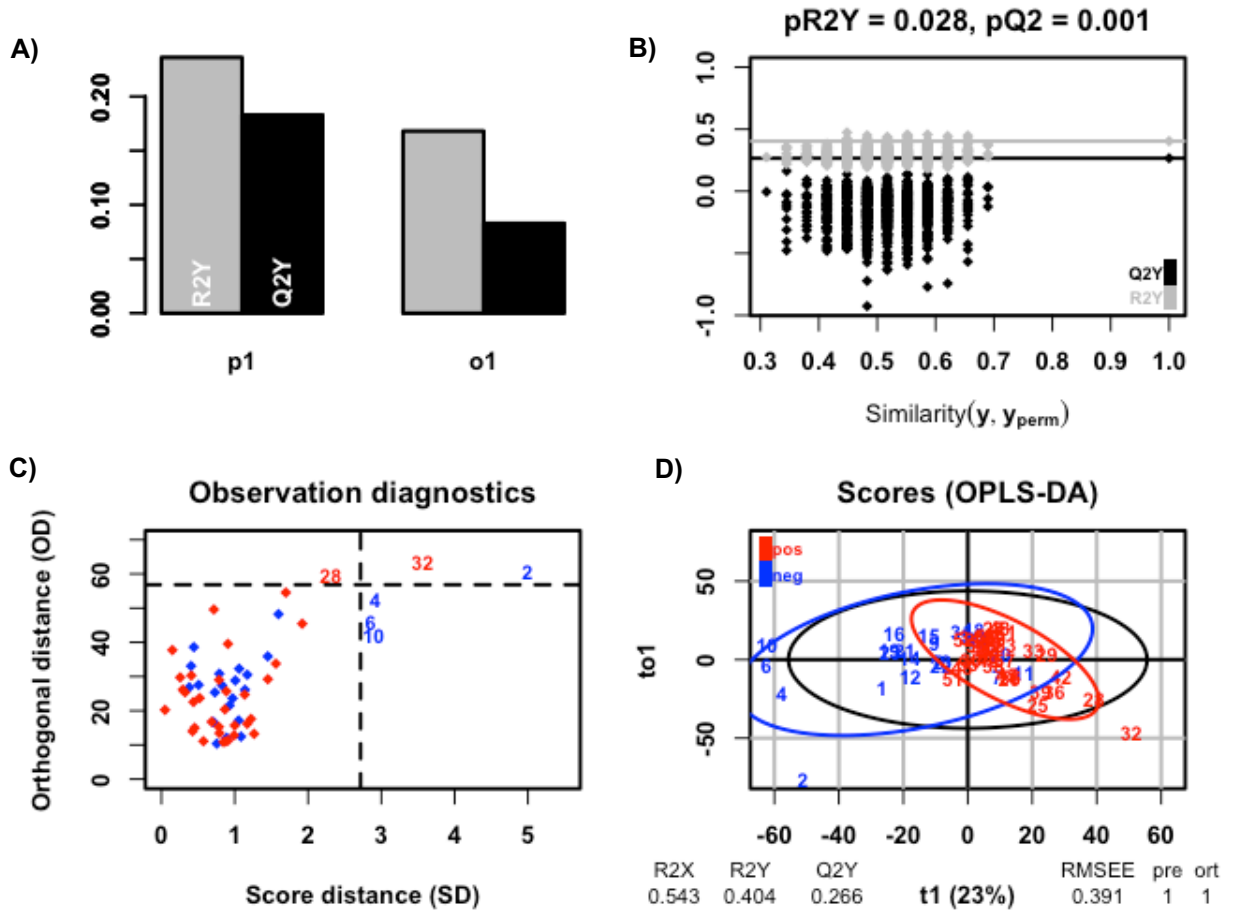
$$TSS = \sum_{n=1}^{I} (y_i - \bar{y})^2 \tag{D16}$$

$$Q^2 = 1 - \frac{PRESS}{TSS} \tag{D17}$$

Onde, SQ_{reg} é a variação explicada pelo modelo, equivalente a $\sum(\hat{y}_i - \bar{y})^2$, SQ_T é a soma quadrática total, $\mathbf{y}^T\mathbf{y}$, SQ_M é a soma quadrática média, y_i é a resposta da i-ésima amostra no vetor $\mathbf{y}$, $\hat{y}_i$ é o i-ésimo valor do vetor resposta $\hat{\mathbf{b}}$ e $\bar{y}$ é o valor médio das observações. Os valores de R^2 e Q^2 são chamados de figuras de mérito, porém outras figuras são, de mesmo modo, importantes para a avaliação da classificação, esses são apresentados no próximo item.

APÊNDICE E – RESULTADOS DO OPLS-DA PARA CONJUNTO DE DADOS DIAGNÓSTICO DE ASMA EM GATOS APÓS DESBALANCEAMENTO.

Figura E1. Resultado do OPLS-DA para o Conjunto de dados para diagnóstico de asma em gatos após o desbalanceamento. A) O R2Y e o Q2Y do modelo. B) O teste de permutação. C) Distâncias para o centro de cada uma das distribuições dos grupos. D) Escores das amostras de treinamento do OPLS-DA, nesse gráfico pos é o grupo 1 e neg é o grupo 0.



Quadro E1. Matriz de confusão do modelo OPLS-DA para o Conjunto de dados para diagnóstico de asma em gatos após o desbalanceamento das classes. E as figuras de mérito para o conjunto de teste.

Modelo OPLS-DA	**Diagnóstico Clínico**		
	0	*1*	*Total*
0	**16(1)**	0(0)	16(1)
1	8(12)	**34(19)**	42(31)
Total	24(13)	34(19)	

Para o conjunto de teste:	
Kappa	0.09
Exatidão	62,5%
Especificidade	7,7%
Sensibilidade	100%

APÊNDICE F – NOTA DE IMPRENSA

Otimização Quimiométrica para Ensaios Metabonômicos Baseados em Espectroscopia de RMN – Diagnóstico e Estadiamento de Câncer de Próstata

(Tese de Doutorado)

Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco
Doutorando: Márcio Felipe de Oliveira (Bolsista CNPq)
Orientador: Prof. Dr. Ricardo Oliveira da Silva

A Metabonômica é uma área do conhecimento que busca desenvolver métodos não-invasivos, ou minimamente invasivos, para o diagnóstico e monitoramento de diferentes doenças. Para isso, são utilizadas amostras de fluidos corporais de fácil acesso, minimizando os riscos aos pacientes, como soro sanguíneo e urina. São obtidos espectros de ressonância magnética nuclear, que são processados usando diferentes ferramentas de estatística multivariada. Uma das principais dificuldades enfrentadas nesse campo de pesquisa e o número pequeno de amostras e com distribuição não-uniforme das amostras nos grupos estudados. Por exemplo: ter mais amostras de voluntários saudáveis que de doentes. A aprendizagem de máquinas possui diversos algoritmos, como os de seleção de variáveis e os de geração de dados artificiais, ou sobreamostragem, que podem ser utilizados para enfrentar esse problema.

Nesta tese de doutorado, foram desenvolvidos modelos metabonômicos para diagnosticar e monitorar câncer de próstata. Foram recrutados, no Ambulatório de Urologia do Hospital das Clínicas da UFPE, 61 voluntários sendo 38 diagnosticados com câncer de próstata e 28 em câncer de próstata. Dos 38 pacientes com câncer, 27 foram monitorados após a prostatectomia. Desses voluntários, 15 apresentaram tumor com baixa probabilidade de ser agressivo (Classificação Gleason ≤ 2), 15 apresentaram alto risco de recidiva (classificação pT ≥ 3 e margem cirúrgica positiva) e 14 apresentaram recidiva bioquímica. Foram construídos modelos para o diagnóstico de câncer próstata, para estimar o grau de agressividade do tumor, o risco de recidiva e para o prognóstico de recidiva bioquímica. Foram testadas 35 combinações usando sete ferramentas de seleção de variáveis e cinco algoritmo de classificação. O modelo para diagnóstico de câncer de próstata apresentou valores de sensibilidade, especificidade e exatidão iguais a 92%, 83% e 88%, respectivamente. O modelo para estadiamento (agressividade do tumor) apresentou sensibilidade, especificidade e exatidão iguais a 75%, 100% e 89%, respectivamente. O modelo para estimar o risco de recidiva apresentou sensibilidade, especificidade e exatidão iguais a 75%, 80% e 78%, respectivamente. Enquanto o modelo de prognóstico de recidiva bioquímica apresentou sensibilidade, especificidade e exatidão iguais a 87,5%, 60% e 80%, respectivamente.

As combinações de algoritmos, que possuem seleção de variáveis, sobreamostragem e algum tipo de regularização no modelo de classificação se sobressaíram frente aos que não usaram essas ferramentas, demonstrando a utilidade do pré-processamento na melhora da classificação em estudos metabonômicos. Além disso, a técnica metabonômica empregada mostra potencial para ser uma alternativa não-invasiva à biopsia prostática, que é muita invasiva, não pode ser repetida em intervalo curto de tempo e exige a ocupação de leito hospitalar. Os resultados da modelagem para o diagnóstico de câncer de próstata foram publicados na revista *Metabolomics* (https://doi.org/10.1007/s11306-023-02067-x).